## Table of Contents

### Introduction & Background

1. Introduction to Proceedings of Workshop on Big Data and Urban Informatics sponsored by National Science Foundation ......................................................... 1  
   *Piyushimita Thakuriah, Nebiyou Tilahun and Moira Zellner*

2. Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery ................................................................. 4  
   *Piyushimita Thakuriah, Nebiyou Tilahun and Moira Zellner*

### Analytics of User-Generated Content

3. Developing an interactive mobile information platform to integrate environmental big data and citizen science in urban environmental management .......................... 33  
   *Yanfu Zhou, Zhenghong Tang, Hongfeng Yu, Tiantian Liu and Yue Gu*

4. Fast food data: Where user-generated content works and where it doesn’t ........ 49  
   *Robert Manduca, Seth Spielman and David Folch*

5. Using Social Media Data to Understand Cities ............................................. 64  
   *Dan Tasse and Jason Hong*

   *Robert Goodspeed and Jacob Yan*

7. Visualizing a City Using Collection of Crowdsensed Sound .......................... 106  
   *Yoshito Tobe, Itaru Usami and Niwat Thepvilojanapong*

8. CyberGIS Enabled Urban Sensing from Volunteered Citizen Participation using Mobile Devices ............................................................................. 123  
   *Junjun Yin, Shaowen Wang and Yizhao Gao*

9. Using Big Data to Study Urban Sentiments: Twitter Data vs. Published Meeting Minutes .......................................................................................... 137  
   *Justin Hollander and Erin Graves*

10. The designer as regulator – design patterns and categorization in citizen feedback systems ......................................................................................... 158  
    *Dietmar Offenhuber*

11. The Potential for Big Data to Improve Neighborhood-Level Census Data ........ 184  
    *Seth Spielman*

### Data behind Urban Big Data

12. Big Data and Survey Research: Supplement or Substitute? ............................. 201  
    *Timothy Johnson and Tom Smith*

13. When Big Data Are Not Complete: A Time Geographic Analysis of Taxicab GPS Traces ................................................................................... 213  
    *Brittany Wood and Mark W. Horner*
14. Considering smartphones: Testing big data veracity via user privacy and trust in location-aware applications ................................................................. 226
   Caitlin Cottrill

15. True lies in big data: Detecting location spoofing based on geographic knowledge .... 253
   Bo Zhao and Daniel Sui

16. Accounting for Heteroscedasticity in Big Data ........................................ 276
   Francisco Antunes, Aidan O’Sullivan, Filipe Rodrigues and Francisco Pereira

Urban Plan-Making

17. Considerations for Using Big Data in Sustainable Plan Making: Citizen Generated Data for Bike-share Planning ......................................................... 294
   Nader Afzalan

18. Planning for the Change: Mapping Sea Level Rise and Storm Inundation in Sherman Island Using 3Di Hydrodynamic Model and LiDAR ....................... 313
   Yang Ju, Wei-Chen Hsu, John D. Radke, William L. Fourt, Wei Lang, Olivier Hoes, Howard Forster, Gregory Biging, Martine Schmidt-Poolman, Rosanna Neuhausler, Amna Alruheil and William F. Maier

19. Parameterizing Land Use Planning: Deploying quantitative analysis methods in the practice of city planning ......................................................... 330
   Talia Kaufmann

20. Estimating driving and parking patterns from cellular network data: An application to plug-in electric vehicle charging infrastructure planning ................. 332
   Jing Dong, Xuesong Zhou and Pengfei Li

Changing Organizational and Educational Perspectives

   Steven French and Camille Barchers

22. Big Spatio-temporal Network Data Analytics for Smart Cities: Research Needs ........ 360
   Venkata M.V. Gunturi, Shashi Shekhar, William Northrop and Andrew Kotz

23. ‘Big’ Data + ‘Open’ Data + ‘Mobile’ Data: A Framework for Urban Informatics .... 379
   Kevin Desouza

24. Emerging Urban Digital Infomediaries and Civic Hacking in an Era of Big Data and Open Data Initiatives ......................................................... 398
   Piyushmita Thakuria, Lise Dirks and Yaye Keita

Urban Knowledge Discovery

25. The Influence of Distance Decay Coefficient on Modularity-based Community Detection in Space-embedded Networks ............................................ 414
   Yaoli Wang and Song Gao

   Jeremy Wu and Rui Zhang
<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>27.</td>
<td>Time-geographic Relationships between Vector Field Densities of Activity Patterns and Transport Systems</td>
<td>Xintao Liu, Wai Yeung Yan and Joseph Chow</td>
</tr>
<tr>
<td>28.</td>
<td>Understanding the Usage of NYC Taxicabs for Vehicle Electrification in a Large Metropolitan Area</td>
<td>Promiti Dutta and Roger Anderson</td>
</tr>
<tr>
<td>29.</td>
<td>The Impact of Land-Use Variables on Free-floating Carsharing Vehicle Rental Choice and Parking Duration</td>
<td>Mubassira Khan, Randy Machemehl and Zhanmin Zhang</td>
</tr>
<tr>
<td>30.</td>
<td>Big urban probe data for the provision of advanced traveler information services and traffic management schemes</td>
<td>Josep Maria Salanova Grau, Evangelos Mitsakis, Iraklis Stamos and Georgia Aifandopoulou</td>
</tr>
<tr>
<td>31.</td>
<td>Identifying Driving Risk Factors to Support Usage-Based Insurance using Smartphone Personalized Driving Data</td>
<td>Yi-Chang Chiu, Yu-Luen Ma and Xianbiao Hu</td>
</tr>
<tr>
<td>32.</td>
<td>Integrating fixed and mobile arterial roadway data sources for transportation analysis</td>
<td>Alexander Bigazzi, Wei Feng, Adam Moore and Miguel Figliozzi</td>
</tr>
<tr>
<td>33.</td>
<td>Inside the Triangle: Does Database Selection Alter our Understanding of Urban Industrial Systems?</td>
<td>Mary Donegan</td>
</tr>
<tr>
<td>34.</td>
<td>The Southern California Housing Bubble: Neighborhood Level Measures and Economic Implications</td>
<td>Johannes Moenius</td>
</tr>
<tr>
<td>35.</td>
<td>Examining Intraurban Migration in the Twin Cities Metropolitan Area using Parcel Data</td>
<td>Shipeng Sun</td>
</tr>
</tbody>
</table>

### Health and Well-Being

<table>
<thead>
<tr>
<th>No.</th>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>37.</td>
<td>UbiActive: A Smartphone-Based Tool for Trip Detection and Travel-Related Physical Activity Assessment</td>
<td>Yingling Fan</td>
</tr>
<tr>
<td>38.</td>
<td>Archive of Many Outdoor Scenes: Using public, online webcams for public health surveillance</td>
<td>Aaron Hipp, Deepi Adlakha, Rebecca Gernes, Agata Kargol and Robert Pless</td>
</tr>
<tr>
<td>39.</td>
<td>Extracting activity patterns from GPS track data</td>
<td>Sungsoon Hwang, Christian Evans and Timothy Hanke</td>
</tr>
<tr>
<td>40.</td>
<td>Big Data, Small Apps: Premises and products of the civic hackathon</td>
<td>Sara Carr and Allison Lassiter</td>
</tr>
</tbody>
</table>
   Li Yin, Qimin Cheng and Zhenfeng Shao

Urban Data Management

42. On City Dashboards and Data Stores ......................................................... 712
    Oliver O’Brien, James Cheshire, Steven Gray, Michael Batty and Andrew Hudson-Smith
43. Centralized Real-Time Platform as a Tool for Urban Studies ......................... 734
    Orlando Cabanas and Celine Steer
44. Supporting Urban Informatics through a Big Data Analytics Online Workbench .... 756
    Christopher Pettit, Robert Stimson, Marcos Nino-Ruiz, Luca Morandini, Ivo Widjaja, 
    Phil Delaney, Martin Tonko, Richard Sinnott, Bill Randolph and Tom Kvan
45. A Big Data Mashing Tool for Measuring Transit System Performance .............. 780
    Gregory Erhardt, Oliver Lock, Elsa Arcaute and Michael Batty
46. Guiding Data-Driven Transportation Decisions ......................................... 808
    Kristin Tuft, Basem Elazzabi, Nathan Hall, Morgan Harvey, Kath Knobe, David 
    Maier and Veronika Megler

Livability & Sustainability

47. Computable Liveability ................................................................. 830
    Glenn Geers
48. Advancing sustainability indicators through text mining: a feasibility demonstration .. 846
    Samuel Rivera, Barbara Minsker, Daniel Work and Dan Roth
49. Combining Physical and Participatory Sensing in Urban Mobility Networks ........ 859
    Xiao-Feng Xie and Zun-Jing Wang
50. Collecting, Mapping, Analyzing, and Archiving Urban Soundscapes ................. 877
    Tae Hong Park

Insights into Social Equity

51. Big Data and Equity in U.S. Cities ....................................................... 896
    Mai Nguyen and Emma Boundy
52. Modeling Social Justice Issues in Urban Zones Using Big Data: Evidence from 
    Bikeshare and Taxicab Data .......................................................... 941
    David King, Jonathan Peters, Adam Davidson and Juan Francisco Saldarriaga
53. On exploring spatial correlations in census derived socioeconomic datasets ......... 962
    Vinay Prabhu, Rohit Negi and Miguel Rodrigues

Emergencies and Crisis Informatics

54. Dynamic Agent Based Simulation of an Urban Disaster using Synthetic Big Data..... 979
    Daniel Felsenstein, A. Yair Grinberger and Michal Lichter
55. Using Social Media to Task Data Collection and Augment Observations in Urban Areas During Emergencies: 2013 Boulder Floods Case Study .................................................. 1013  
   Guido Cervone, Emily Schnebele, Nigel Waters, Jeff Harrison, Martina Moccaldi and Rosa Sicignano

56. Agent-based Large-Scale Emergency Evacuation Using Real-Time Open Government Data ................................................................. 1027  
   Wei Lu, Cheng Liu and Budhendra Bhaduri

   Rajasekar Karthik and Wei Lu

**Urban Knowledge Discovery  Transportation**

58. From Global to Local: Big Data and Model Development for Spatially and Temporally Scalable Freight Transportation ................................................. 1075  
   Guoqiang Shen

59. Big Brother is Watching You... To Predict Crashes ........................................ 1099  
   Paul St-Aubin, Nicolas Saunier and Luis Miranda-Moreno

60. A Bayesian Urban Link Travel Time Estimation Model Using Large-scale Taxi Trip Data ......................................................................................... 1129  
   Xianyuan Zhan and Satish Ukkusuri

61. Modeling Taxi Demand and Supply in New York Using Large-Scale Taxi GPS Data .. 1149  
   Ci Yang and Eric Gonzales

62. Finding Public Transportation Community Structure based on Large-Scale Smart Card Records in Beijing ......................................................... 1173  
   Song Gao and Ying Long

63. Spatially-Explicit Computational Evaluation of Urban Accessibility from a Human’s Point of View ................................................................. 1188  
   Eran Ben-Elia, Itzhak Benenson, Yodan Rofe and Dimitry Geyzersky

64. Developing a Comprehensive US Transit Accessibility Database .......................... 1209  
   Andrew Owen and David Levinson

65. Impact of Subway Accessibility on Taxi Trip Generation in New York City .......... 1229  
   Dan Wan, Camille Kamga and Anil Yazici

**Urban Knowledge Discovery  Cities, Land Use, Energy**

66. Exploring the Use of Big Data’ for Analyzing the Dynamic Spatial Patterns of Hotel and Entertainment Development ................................................. 1327  
   Li Yin, Liang Wu, Sam Cole and Laiyun Wu

67. Popular Urban Amenities: The ingredients of attractive restaurant areas in Dutch cities ......................................................................................... 1258  
   Daniel Arribas-Bel and Jessie Bakens

68. Getting out of the pool: shrinkage methods for tall and wide datasets. An example from building energy consumption .......................................... 1275  
   David Hsu
   Hossein Estiri

Appendix & Index

70. Appendix A. Program Committee ..................................................... 1324

71. Appendix B. Workshop I Participants ............................................. 1325

71. Author Index ............................................................. 1326

72. Keyword Index............................................................. 1331
Introduction to Proceedings of Workshop on Big Data and Urban Informatics sponsored by National Science Foundation

Piyushimita (Vonu) Thakuriah, University of Glasgow, Workshop Chair
Nebiyou Yonas Tilahun, University of Illinois at Chicago, Workshop Co-Chair
Moira Zellner, University of Illinois at Chicago, Workshop Co-Chair

The Workshop on Big Data and Urban Informatics held at the University of Illinois at Chicago on August 11 and 12, 2014 and sponsored by the National Science Foundation provided a unique opportunity to bring together urban social scientists and data scientists interested in the use of Big Data to address urban challenges. A background paper titled “Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery” documenting the major motivations for the workshop is given in Chapter 2 of these proceedings and is also available online at https://urbanbigdata.uic.edu/files/2015/11/Background.pdf.

There has been a lot of buzz about Big Data, and the term has been used to mean different types or attributes of data. In the urban sector, Big Data is generated in different ways, allowing an exploration and analysis of patterns and dynamics underpinned by complex urban challenges that cannot be studied only by using data from traditional surveys, census or other conventional data systems. We use Big Data to mean data that are generated through transactional, operational, planning and social activities, which are not specifically designed for research, or the linkage of such data to traditional data. The resultant complexities associated with the generation, access and use of such data require special considerations for technology and information management, methods of analysis, and the underlying scientific paradigm and political economy supporting inquiry. Big Data provides significant opportunities for the advancement of Urban Informatics, which is the exploration and understanding of urban patterns and dynamics undertaken from both a theory-driven perspective, as well as an empirical data-driven perspective for conceptual insights and knowledge discovery regarding urban systems, and for improved urban operations, planning and policy, citizen engagement, governance, and innovations.

Urban Big Data can be generated through sensor systems in urban infrastructure and moving objects such as cars. It can be user-generated through social media, human computation tasks and citizen science projects. It can refer to administrative data accessed through open data portals or through restricted access to highly confidential governmental records, as well as to private sector data on consumer transactions on goods and service agreements. It can also be highly unstructured data from arts and digital humanities collections. Additionally, it encompasses hybrid data sources such as linked survey and sensor data. There are now significant user communities with interest in these types of urban data sources spanning across a range of disciplines, as well as private industry and governments.

We organized an initial workshop, which took place on March 28th and 29th, 2013 in the University of Illinois at Chicago, to help generate ideas for a larger event. Twenty-one academics and practitioners and three PhD students (listed in Appendix A) participated in this workshop, representing 9 institutions in the US, UK and Australia. The disciplines represented were Urban Planning, Computer Science, Civil Engineering, Economics, Statistics, and Geography. The participants of the first workshop were identified during the proposal-writing phase and served on the program committee of the main workshop, and helped to delineate the purpose driving this event.
Our main motivation for the main workshop, which took place in August 2014, was to convene researchers and professionals working on this interdisciplinary topic, and to organize a community with interests on theoretical developments and applications demonstrating the use of urban Big Data, and the next-generation of Big Data services, tools and technologies for Urban Informatics. We were interested in research results as well as idea pieces and works in progress that highlighted research needs and data limitations. We sought papers that clearly create or use novel, emerging sources of Big Data for urban and regional analysis in transportation, environment, public health, land-use, housing, economic development, labor markets, criminal justice, population demographics, urban ecology, energy, community development, and public participation.

The scope of the workshop ranged across five key strands of the Big Data debate in the urban context:

*Theoretical developments and knowledge discovery in urban systems*: Theoretical developments, knowledge discovery and hypothesis generation about urban dynamics and processes, and understanding of urban systems and their social, behavioral, political, mobility and economic aspects including models of transactions, incentives, collaboration, and cooperation, and behavioral or organizational change, as well as epistemological positions and scientific paradigms regarding modes of inquiry, and investigations into the validity of approaches used and limits of knowledge about urban systems derived from datafication;

*Information management for Urban Informatics*: Information management for Urban Informatics including solutions to managing urban information such as data infrastructure management solutions, search and querying, data security and privacy, resource discovery, and language and execution environments, and computational approaches relating to information gathering, management and distribution of urban Big Data;

*Measurement, analysis and methodological research*: Approaches for urban Big Data information retrieval, extraction, linkage, analytics and visualization; advancements in exploratory and predictive analytics; agent-based modeling and complex systems; statistical inference, data quality and related issues such as missing data, endogeneity and selection biases; assessment of the extent to which urban Big Data may be able to add to traditional survey-based urban social science research, as well as examples of Big Data linkage with census, survey and administrative data; evolving goodness of fit metrics for models and data; methods to handle verification and validation, and sensitivity analysis; and experimental design;

*Planning, policy analysis and operational uses of urban Big Data*: Empirical research exemplifying Big Data for improved planning, management and governance in the urban sectors including uses for decision-making, and development of indicators to monitor economic and social activity, and for urban sustainability, transparency, livability, social inclusion, place-making, accessibility and resilience;

*Institutional issues, organizations, networks and infomediaries in urban Big Data*: Studies of locational privacy, trust management and information security relating to urban Big Data; analysis of social networks and sensing systems involved in urban data; studies of civic hacking networks, new types of digital entrepreneurship, organizational assessments of open data portals and city dashboards, questions of political economy relating to information networks and power structures, data access, data confidentiality and security, data governance framework, provenance and ethics.
The main workshop on Big Data and Urban Informatics took place in Aug 11-12, 2014 at the University of Illinois at Chicago and was attended by approximately 150 researchers, educators, practitioners and students from urban social sciences and data sciences interested in the use of Big Data to address urban challenges. Participants were from 11 countries – USA, UK, China, Canada, Japan, Portugal, Greece, Ireland, Israel, Italy, and Australia – representing 91 different institutions. We received 91 initial submissions of which a total of 68 papers were presented over a course of two days, ranging across the five themes.

Papers presented can be classified into 12 broad categories as follows:

1. Analytics of User-Generated Content (9 papers)
2. Data behind Urban Big Data (5 papers)
3. Urban Plan-Making (4 papers)
4. Changing Organizational and Educational Perspectives with Big and Open Urban Data (4 papers)
5. Urban Knowledge Discovery- General (12 papers)
6. Urban Knowledge Discovery – Transportation (8 papers)
7. Urban Knowledge Discovery – Cities, Land Use, Energy (4 papers)
8. Health and Well-Being (5 papers)
9. Urban Data Management (5 papers)
10. Livability & Sustainability (4 papers)
11. Insights into Social Equity (4 papers)
12. Emergencies and Crisis Informatics (4 papers)

Selected papers from the workshop will be published, after additional peer-review, in an edited volume titled *Seeing Cities Through Big Data - Research, Methods and Applications in Urban Informatics* anticipated to be released in early 2016. Papers presented at the workshop have also been released through the workshop’s website at http://urbanbigdata.uic.edu/proceedings and are publicly available for researchers. The research team also used social media and academic and non-academic mailing lists to announce both the conference as well as the release of the conference proceedings.

We wish to thank the National Science Foundation and the University of Illinois at Chicago’s Department of Urban Planning and Policy for making this workshop possible. We also thank the authors for their contribution, the program committee for their reviews, and the workshop attendees for an engaging two-day event. A very special thanks to Nina Savar, workshop coordinator, without whose diligent work the workshop would not have been possible.
Big Data and Urban Informatics: Innovations and Challenges to Urban Planning and Knowledge Discovery

Piyushimita (Vonu) Thakuriah, University of Glasgow
Nebiyou Tilahun, University of Illinois at Chicago
Moira Zellner, University of Illinois at Chicago

Please cite as:

Forthcoming as:

Abstract

Big Data is the term being used to describe a wide spectrum of observational or “naturally-occurring” data generated through transactional, operational, planning and social activities that are not specifically designed for research. Due to the structure and access conditions associated with such data, research and analysis using such data becomes significantly complicated. New sources of Big Data are rapidly emerging as a result of technological, institutional, social, and business innovations. The objective of this background paper is to describe emerging sources of Big Data, their use in urban research, and the challenges that arise with their use. To a certain extent, Big Data in the urban context has become narrowly associated with sensor (e.g., Internet of Things) or socially generated (e.g., social media or citizen science) data. However, there are many other sources of observational data that are meaningful to different groups of urban researchers and user communities. Examples include privately held transactions data, confidential administrative micro-data, data from arts and humanities collections, and hybrid data consisting of synthetic or linked data.

The emerging area of Urban Informatics focuses on the exploration and understanding of urban systems by leveraging novel sources of data. The major potential of Urban Informatics research and applications is in four areas: (1) improved strategies for dynamic urban resource management, (2) theoretical insights and knowledge discovery of urban patterns and processes, (3) strategies for urban engagement and civic participation, and (4) innovations in urban management, and planning and policy analysis. Urban Informatics utilizes urban Big Data in innovative ways by retrofitting or repurposing existing urban models and simulations that are underpinned by a wide range of theoretical traditions, as well as through data-driven modeling approaches that are largely theory agnostic, although these divergent research approaches are starting to converge in some ways. The paper surveys the kinds of urban problems being considered by going from a data-poor environment to a data-rich world and ways in which such enquiry have the potential to enhance our understanding, not only of urban systems and processes overall, but also contextual peculiarities and local experiences. The paper concludes by commenting on challenges that are likely to arise in varying degrees when using Big Data for Urban Informatics: technological, methodological, theoretical/epistemological, and the emerging political economy of Big Data.

Keywords: Big Data, Urban Informatics, Knowledge Discovery, dynamic resource management, user-generated content

1. Introduction

Urban and regional analysis involve the use of a wide range of approaches to understand and manage complex sectors, such as transportation, environment, health, housing, the built environment, and the economy. The goals of urban research are many, and include theoretical understanding of infrastructural, physical and socioeconomic systems; developing approaches to improve urban operations and management; long-range plan making, and impact assessments of urban policy.
Globally, more people live in urban areas than in rural areas, with 54% of the world’s population estimated to be residing in urban areas in 2014 (United Nations, 2014), levying unprecedented demand for resources and leading to significant concerns for urban management. Decision-makers face a myriad of questions as a result, including: What strategies are needed to operate cities effectively and efficiently? How can we evaluate potential consequences of complex social policy change? What makes the economy resilient and strong and how do we develop shockproof cities? How do different cities recover from man-made or natural disasters? What are the technological, social and policy mechanisms needed to develop interventions for healthy and sustainable behavior? What strategies are needed for lifelong learning, civic engagement, and community participation, adaptation and innovation? How can we generate hypothesis about the historical evolution of social exclusion and the role of agents, policies and practices?

The Big Data tsunami has hit the urban research disciplines just like many other disciplines. It has also stimulated the interest of practitioners and decision-makers seeking solutions for governance, planning and operations of multiple urban sectors. The objective of this background paper is to survey the use of Big Data in the urban context across different academic and professional communities, with a particular focus on Urban Informatics. Urban Informatics is the exploration and understanding of urban systems for dynamic resource management, knowledge discovery and understanding of urban patterns and dynamics, urban engagement and civic participation, and urban planning and policy analysis. Urban Informatics research approaches involve both a theory-driven as well as an empirical data-driven perspective centered on emerging Big Data sources. New sources of such data are arising as a result of technological, institutional, social and business innovations, dramatically increasing possibilities for urban researchers. Equally importantly, new ways of accessing existing sources of data, or innovations in the linkage of data belonging to different owners and domains are leading to new connected data systems. We identify major research questions that may be possible to investigate with the data, as well as existing questions that can be revisited with improved data, in an attempt to identify broad themes for the use of Big Data in Urban Informatics.

While the main research agenda is about better understanding and knowledge discovery of urban systems, there are equally important questions relating to technical challenges in managing the data and in addressing methodological and measurement questions that arise. The use of Big Data in Urban Informatics pose significant epistemological challenges regarding the overall modes of research inquiry, and about institutions and the overall political economy regarding the access and use.

This chapter is organized as follows: in Section 2, we review background information and different types of Big Data being used for urban research. This is followed in Section 3 by a discussion of research approaches and applications in Urban Informatics that involve the use of Big Data. Challenges that arise with the use of such data are discussed next in Section 4 and conclusions are drawn in Section 5.

2. Big Data: Complexities and Types

For many, Big Data is just a buzzword and to a certain extent, the ambiguity in its meaning reflects the different ways in which it is used in different disciplines and user communities. The ambiguity is further perpetuated by the multiple concepts that have become associated with the topic. However, the vagueness and well-worn clichés surrounding the subject have overshadowed potentially strong benefits in well-considered cases of use.

Based on a review of 1,437 conference papers and articles that contained the full term “Big Data” in either the title or within the author-provided keywords, De Mauro et al. (2014) arrived at four groups of definitions of Big Data. These definitions focus on: (1) the characteristics of Big Data (massive, rapid, complex, unstructured and so on), with the 3-Vs - Volume, Variety and Velocity - referring to the pure amount of information and challenges it poses (Laney, 2001) being a particularly over-hyped
example; (2) the technological needs behind the processing of large amounts of data (e.g., as needing
serious computing power, or, scalable architecture for efficient storage, manipulation, and analysis);
(3) as Big Data being associated with crossing of some sort of threshold (e.g., exceeding the
processing capacity of conventional database systems); and (4) as highlighting the impact of Big Data
advancement on society (e.g., shifts in the way we analyze information that transform how we
understand and organize society).

Moreover, the term Big Data has also come to be associated with not just the data itself, but with
curiosity and goal-driven approaches to extract information out of the data (Davenport and Patil,
2012), with a focus on the automation of the entire scientific process, from data capture to processing
to modeling (Pietsch, 2013). This is partly an outcome of the close association between Big Data and
data science, which emphasizes data-driven modeling, hypothesis generation and data description in a
visually appealing manner. These are elements of what has become known as the Fourth Paradigm of
scientific discovery (Gray, 2007 as given in Hey et al., 2009), which focuses on exploratory, data-
intensive research, in contrast to earlier research paradigms focusing on describing, theory-building
and computationally simulating observed phenomenon.

Quantitative urban research has historically relied on data from censuses, surveys, and specialized
sensor systems. While these sources of data will continue to play a vital role in urban analysis,
declining response rates to traditional surveys, and increasing costs of administering the decennial
census and maintaining and replacing sensor systems have led to significant challenges to having
high-quality data for urban research, planning and operations. These challenges have led to
increasing interest in looking at alternative ways of supplementing the urban data infrastructure.

For our purposes, Big Data refers to structured and unstructured data generated naturally as a part
transactional, operational, planning and social activities, or the linkage of such data to purposefully
designed data. The use of such data gives rise to technological and methodological challenges and
complexities regarding the scientific paradigm and political economy supporting inquiry. Established
and emerging sources of urban Big Data are summarized in Table 1: sensor systems, user-generated
content, administrative data (open and confidential micro-data), private sector transactions data, data
from arts and humanities collections, and hybrid data sources, including linked data and synthetic
data. While there are many ways to organize Big Data for urban research and applications, the
grouping here is primarily informed by the user community typically associated with each type of
data; other factors such as methods of generation, and issues of ownership and access are also
considered. The grouping is not mutually exclusive; for example, sensor systems might be owned by
public agencies for administrative and operational purposes as well as by private companies to assist
with transactions.
Table 1: Types of Urban Big Data and Illustrative User Communities

<table>
<thead>
<tr>
<th>Urban Big Data</th>
<th>Examples</th>
<th>Illustrative User Communities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensor systems</td>
<td>Environmental, water, transportation, building management sensor systems; connected systems; Internet of Things</td>
<td>Public and private urban operations and management organizations, independent ICT developers, researchers in the engineering sciences</td>
</tr>
<tr>
<td>User-Generated Content (&quot;social&quot; or &quot;human&quot; sensors)</td>
<td>Participatory sensing systems, citizen science projects, social media, web use, GPS, online social networks and other socially generated data</td>
<td>Private businesses, customer/client-focused public organizations, independent developers, researchers in data sciences and urban social sciences</td>
</tr>
<tr>
<td>Administrative (governmental) data (open and confidential micro-data)</td>
<td>Open administrative data on transactions, taxes and revenue, payments and registrations; confidential person-level micro-data on employment, health, welfare payments, education records</td>
<td>Open data: innovators, civic hackers, researchers Confidential data: government data agencies, urban social scientists involved in economic and social policy research, public health and medical researchers</td>
</tr>
<tr>
<td>Private Sector Data (customer and transactions records)</td>
<td>Customer transactions data from store cards and business records; fleet management systems; customer profile data from application forms; usage data from utilities and financial institutions; product purchases and terms of service agreements</td>
<td>Private businesses, public agencies, independent developers, researchers in data sciences and urban social sciences</td>
</tr>
<tr>
<td>Arts and Humanities Data</td>
<td>Repositories of text, images, sound recordings, linguistic data, film, art and material culture, and digital objects, and other media</td>
<td>Urban design community, historical, art, architecture and digital humanities organizations, community organizations, data scientists and developers, private organizations</td>
</tr>
<tr>
<td>Hybrid data (linked and synthetic data)</td>
<td>Linked data including survey-sensor, census-administrative records</td>
<td>Urban planning and social policy community, government data organizations, private businesses and consultants</td>
</tr>
</tbody>
</table>

2.2.1 Sensor Systems: Infrastructure and Moving Object Sensors and Internet of Things

Sensors in urban infrastructure (transportation, health, energy, water, waste, weather systems, structural health monitoring systems, environmental management, buildings and so on) result in vast amounts of data on urban systems. Novel patterns of demand and usage patterns can be extracted from these data. The sensors detect activity and changes in a wide variety of urban phenomena involving
inanimate objects (infrastructure, building structure), physical aspects of urban areas (land cover, water, tree cover, and atmospheric conditions), movement (of cars, people, animals), and activity (use patterns, locations).

As noted earlier, sensor systems might be government or privately owned, with very different access and data governance conditions, and some have been operational for a long time. Typical user communities are public and private urban operations management organizations, independent ICT developers, and researchers in the engineering sciences. However, sensor data, if linked to other sources and archived over long periods of time, can be used by urban social scientists studying long-term social, economic and environmental changes, and their effects on neighborhoods and communities. The emerging smart cities community has become increasingly involved with urban sensor systems, particularly with their integration and performance enhancement through ICT solutions. Many urban sensor systems are now likely to be wirelessly connected, mobile, and significantly more embedded and distributed. Examples from a vast range of operational and planned applications include cooperative or connected vehicle systems, Vehicle-to-Grid systems, Smart Grid systems, and a wide range of indoor and outdoor assistive technologies for seniors and persons with disabilities. The diverse field of remote sensing has been undergoing rapid developments as well, with massive amounts of high-resolution temporal and spatial data being collected more rapidly than ever before with sensors that are mounted on satellites, planes, and lately, drones.

Potentially the “next phase” in this ever-transforming technology landscape is the creation of tiny, intelligent devices that are embedded into everyday objects such as houses, cars, furniture, and clothes, and which can “listen in” and produce recommendations and interventions as needed. The concept of the “Internet of Things” (IoT) attributed to Ashton (1999) is still primarily a vision at this stage, although there are many individual IoT technologies and systems that are operational, although a future with Machine-to-Machine (M2M) communications is envisioned by some, where “billions to trillions of everyday objects and the surrounding environment are connected and managed through a range of devices, communication networks, and cloud-based servers” (Wu, 2011). Needless to say, the number and variety of data streams available to study cities will greatly increase.

2.2.2 User-Generated Content: Social and Human Sensing Systems

Transformative changes have taken place in the last decade regarding ways in which citizens are being involved in co-creating information, and much has been written about crowd-sourcing, Volunteered Geographic Information, and, generally, User-Generated Content (UGC). Citizens, through the use of sensors or social media, and other socially generated information resulting from their participation in social, economic or civic activities, are going from being passive subjects of survey and research studies to being active generators of information. Citizen-based approaches can be categorized as contributory, collaborative, or co-created (Bonney et al., 2009). UGC can generally occur: (1) proactively when users voluntarily generate data on ideas, solve problems, and report on events, disruptions or activities that are of social and civic interest, or (2) retroactively, when analysts process secondary sources of user-submitted data published through the web, social media and other tools (Thakuriah and Geers, 2013).

UGC can be proactively generated through idea generation, feedback and problem solving. Developments in Information and Communications Technology (ICT) have expanded the range and diversity of ways in which citizens provide input into urban planning and design sourcing, vote on and share ideas about urban projects, and provide feedback regarding plans and proposals with the potential to affect life in cities. These range from specialized focus groups where citizens provide input to “hackathons” where individuals passionate about ICT and cities get together to generate solutions to civic problems using data. Citizens also solve problems; for example, through human computation (described further in Section 3.4) to assess livability or the quality of urban spaces where objective metrics from sensors and machines are not accurate. These activities produce large volumes
of structured and unstructured data that can be analyzed to obtain insights into preferences, behaviors and so on.

There has been an explosive growth in the wealth of data proactively generated through different sensing systems. Depending on the level of decision-making needed on the part of users generating information, proactive sensing modes can be disaggregated into participatory (active) and opportunistic (passive) sensing modes. In participatory sensing, users voluntarily opt into systems that are specifically designed to collect information of interest (e.g., through apps which capture information on quality of local social, retail and commercial services, or websites that consolidate information for local ride-sharing), and actively report or upload information on objects of interest. In opportunistic sensing, users enable their wearable or in-vehicle location-aware devices to automatically track and passively transmit their physical sensations, or activities and movements (e.g., real-time automotive tracking applications which measure vehicle movements yielding data on speeds, congestion, incidents and the like, as well as biometric sensors, life loggers and a wide variety of other devices for personal informatics relating to health and well-being). The result of these sensing programs are streams of content including text, images, video, sound, GPS trajectories, physiological signals and others, which are available to researchers at varying levels of granularity depending on, among other factors, the need to protect personally identifiable information.

In terms of retroactive UCG, massive volumes of content are also created every second of every day as a result of users providing information online about their lives and their experiences. The key difference from the proactive modes is that users are not voluntarily opting into specific systems for the purpose of sharing information on particular topics and issues. There are many different types of retroactive UGC that could be used for urban research including Internet search terms, customer ratings, web usage data, and trends data. Data from social networks, micro-blogs or social media streams have generated a lot of interest among researchers, with the dominant services at the current time being online social networks such as Twitter, Facebook and LinkedIn, and Foursquare, the latter being a location-based social network. Additionally, there are general question-and-answer databases from which data relevant to urban researchers could be retrieved, as well as a wide variety of multimedia online social sharing platforms such as YouTube and Flickr, and user-created online content sites such Wikipedia, TripAdvisor and Yelp. Such naturally occurring UGC provide a rich source of secondary data on the social fabric of cities, albeit through the lens of their user communities, raising questions regarding bias and lack of generalizability. However, provided appropriate information retrieval and analytics techniques are used, such data can allow detection and monitoring of events and patterns of interest, as well as the ability to identify concerns, emotions and preferences among citizens, particularly in response to news, urban operation disruptions and policy changes, for real-time understanding of urban dynamics.

### 2.2.3 Administrative Data

Governments collect micro-data on citizens as a part of their everyday business or operational processes on registration, transactions and record keeping which typically occur during the delivery of a service. Tax and revenue agencies record data on citizens and taxes paid revenues generated, licenses issued and real estate or vehicle transactions made. Employment and benefits agencies collect information on income, earnings and disability or retirement benefits. Administrative micro-data in particular contain a wealth of information that is relevant to urban policy evaluation. The advantages often cited regarding the use of administrative data in research include being relatively cheap and potentially less intrusive and yet comprehensive (Gowans et al., 2015), as well as having larger sample sizes, and fewer problems with attrition, non-response, and measurement error compared to traditional survey data sources (Card et al., n.d.).
One particular development with administrative data is the increasing availability of administrative and other governmental data through “Open Data” initiatives. These initiatives have largely been driven by open government strategies, generally thought of as reflecting transparent government, collaborative government, and innovative government, with some degree of confusion and ideological tensions about what these terms mean in practice (Shkabatur, 2013; Pyrozhenko, 2011). Open data initiatives are based on the idea that governmental data should be accessible for everyone to use and to republish without copyright or other restrictions in order to create a knowledgeable, engaged, creative citizenry, while also bringing about accountability and transparency. Open Data initiatives have the potential to lead to innovations (Thorhildur et al., 2013) and to address the needs of the disadvantaged (Gurstein, 2011).

National and local governments around the world have now supported open data policies. This has led to a proliferation of open data portals where government agencies upload administrative data that are aggregated or anonymized by removing personal identifiers, and is license-free and in non-proprietary formats. Although they present many opportunities, open data initiatives can face challenges due to a number of reasons including closed government culture in some localities, privacy legislation, limitations in data quality that prohibit publication, and limited user-friendliness (Huijboom and van den Broek, 2011).

Many valuable uses of administrative data require access to personally identifiable information, typically micro-data at the level of individual persons, which are usually strictly protected by data protection laws or other governance mechanisms. Personally-identifiable information are those that can be used on its own or together with other information to identify a specific individual, and the benefits of accessing and sharing identifiable administrative data for research purposes have to be balanced against the requirements for data security to ensure the protection of individuals’ personal information. Confidential administrative micro-data are of great interest to urban social scientists involved in economic and social policy research, as well as to public health and medical researchers.

There are several activities currently that are likely to be of interest to urban researchers. The UK Economic and Social Research Council recently funded four large centers on administrative data research, including running data services to support confidential administrative data linkage, in a manner similar to that offered in other countries such as Denmark, Finland, Norway and Sweden. In the US, the Longitudinal Employment Household Dynamics (LEHD) program of the Census Bureau is an example of an ambitious nationwide program combining federal, state and Census Bureau data on employers and employees from unemployment insurance records, data on employment and wages, additional administrative data and data from censuses and surveys (Abowd et al., 2005), to create detailed estimates of workforce and labor market dynamics.

Administrative data in some cases can be linked both longitudinally for the same person over time and between registers of different types, e.g. linking employment data of parents to children’s test scores, or linking medical records to person’s historical location data and other environmental data. The latter, for example, could potentially allow research to investigate questions relating to epigenetics and disease heritability (following Aguilera et al., 2010). Such linkages are also likely to allow in-depth exploration of spatial and temporal variations in health and social exclusion.

2.2.4 Private Sector Transaction Data

Like government agencies, businesses collect data as a part of their everyday transactions with customers. They also develop detailed customer profiles from different sources. Such privately held data may be contrasted with the aforementioned privately owned sensor systems data as those that continuously track customer activity and use patterns. In a report titled “New Data for Understanding the Human Condition: International Perspectives” (OECD Global Science Forum, 2013), customer transactions was identified as a major data category, within which the following were noted as useful
data sources: store cards such as supermarket loyalty cards, customer accounts on utilities, financial institutions, and other customer records such as product purchases and service agreements.

Companies have historically used such data to improve business process management, market forecasts and to improve customer relations. Many of these data sources can provide key insights into challenges facing cities and have been increasingly of interest to urban researchers. For example, utility companies have records on energy consumption and transactions, which can help to understand variations in energy demand and impact for sustainable development policy, while also understanding implications for fuel poverty where households spend more than some acceptable threshold to maintain adequate heating (NAREC, 2013).

2.2.5 Arts and Humanities Collections and Historical Urban Data

There are vast arts and humanities collections that depict life in the city that include text, image, sound recording, and linguistic collections, as well as media repositories such as film, art, material culture, and digital objects. These highly unstructured sources of data allow the representation of the ocular, acoustic and other patterns and transformations in cities to be mapped and visualized, to potentially shed light on social, cultural and built environment patterns in cities. For example, a recent project seeks to digitize a treasure trove of everyday objects such as “advertisements, handbills, pamphlets, menus, invitations, medals, pins, buttons, badges, three-dimensional souvenirs and printed textiles, such as ribbons and sashes” to provide “visual and material insight into New Yorkers’ engagement with the social, creative, civic, political, and physical dynamics of the city, from the Colonial era to the present day” (Museum of the City of New York, 2014), which will have detailed metadata making it searchable through geographic querying.

Inferring knowledge from such data involves digitization, exploratory media analysis, text and cultural landscape mapping, 3-D mapping, electronic literary analysis, and advanced visualization techniques. With online publishing and virtual archives, content creators and users have the potential to interact with source materials to create new findings, while also facilitating civic engagement, community building and information sharing. Recent focus has been on humanities to foster civic engagement; for example, the American Academy of Arts and Sciences (2013), while making a case for federal funding for the public humanities, emphasized the need to encourage “civic vigor” and to prepare citizens to be “voters, jurors, and consumers”. There is potential for this line of work in improving the well-being of cities by going beyond civic engagement, for example, to lifelong learning (Hoadley and Bell, 1996; CERI/OECD, 1992). Stakeholders engaged in this area are typically organizations involved in cultural heritage and digital culture, such as museums, galleries, memory institutions, libraries, archives and institutions of learning. Typical user communities for this type of data are history, urban design, art and architecture, and digital humanities organizations, as well as community and civic organizations, data scientists, and private organizations. The use of such data in quantitative urban modeling opens up a whole new direction of urban research.

2.2.6 Hybrid Data and Linked Data Systems

Data combinations can occur in two ways: combination through study design to collect structured and unstructured data during the same data collection effort (e.g., obtaining GPS data from social survey participants, so that detailed movement data are collected from the persons for whom survey responses are available), or through a combination of different data sources brought together data by data linkage or multi-sensor data fusion under the overall banner of what has recently been called “broad data” (Hendler, 2014).

There are now several examples where data streams have been linked by design, an example of which is household travel surveys and activity diaries have been administered using both questionnaire-based survey instrument and a GPS element. One of many examples is the 2007/2008 Travel Tracker data collection by the Chicago Metropolitan Agency for Planning (CMAP), which included travel
diaries collected via computer assisted telephone interviews (CATI) and GPS data collected from a subset of participants over 7 days. Recent efforts have expanded the number of sensing devices used and the types of contextual data collected during the survey period. For example, the Integrated Multimedia City Data (iMCD) (Urban Big Data Center, n.d.), which is being administered at the time of writing this paper, involves a questionnaire-based survey covering travel, ICT use, education and literacy, civic and community engagement, and sustainable behavior of a random sample of households in Glasgow, UK. Respondents undertake a sensing survey using GPS and life logging sensors leading to location and mobility data and rapid still images of the world as the survey respondent sees it. In the survey background is a significant Information Retrieval effort from numerous social media and multimedia web sources, as well as retrieval of information from transport, weather, crime-monitoring CCTV and other urban sectors. Alongside these data streams are Very High Resolution satellite data and LiDAR allowing digital surface modeling creating 3D urban representations.

The census is the backbone for many types of urban analysis; however, its escalating costs has been noted to be unsustainable, with the cost of the 2010 US Census being almost $94 per housing unit, representing a 34% increase in the cost per housing unit over Census 2000 costs, which in turn represents a 76% increase over the costs of the 1990 Census (Reist and Ciango, 2013). There was an estimated net undercount of 2.07% for Blacks, 1.54% for Hispanics, and 4.88% for American Indians and Alaska Natives, while non-Hispanic whites had a net over-count of 0.83 percent (Williams, 2012). Vitrano and Chapin (2014) estimated that without significant intervention, the 2020 Census would cost about $151 per household. This has led the US Census Bureau to actively consider innovative solutions designed to reduce costs while maintaining a high quality census in 2020. Some of the strategies being considered include leveraging the Internet and new methods of communications to improve self-response by driving respondents to the Internet and taking advantage of Internet response processes. Another census hybridization step being considered is the use of administrative records to reduce or eliminate some interviews of households that do not respond to the census and related field contacts.

Similar concerns in the UK led to the Beyond 2011 program where different approaches to produce population statistics were considered. The program recommended several potential approaches such as the use of an online survey for the decennial census and a census using existing government data and annual compulsory surveys (Office for National Statistics, 2015). The ONS Big Data project is also evaluating through a series of pilot projects the possibility of using web scraping of Internet price data for the Consumer Price Index (CPI) and the Retail Price Index (RPI) and Twitter data to infer student movement, which is a population that has been historically been difficult to capture through traditional surveys (Naylor et al., 2015). Other Big Data sources being studied as a part of the pilots are smart meter data to identify household size/structure and the likelihood of occupancy during the day, and mobile phone positioning data to infer travel patterns of workers.

Another situation is where data on survey respondents are linked to routine administrative records; one approach involved the use of an informed consent process where respondents who agree to participate in a survey are explicitly asked if the information they provide can be linked to their administrative records. One example of this approach is the UK Biobank Survey (Lightfoot and Dibben, 2013). Having survey responses linked to administrative data enables important urban policy questions to be evaluated; the key issue here is that participants understand and agree to such linkage.

From urban operations point of view, connected systems allow a degree of sophistication and efficiency not possible with data from individual data systems. This was touched upon briefly in Section 2.2.1; clearly weather-responsive traffic management systems (Thakuriah and Tilahun, 2013) and emergency response systems (Salaszyk et al., 2006) require extensive integration of very different streams of data, often in real time. This can be computationally challenging, but also perhaps
equally challenging to get data owners to share information. These types of linked data are likely to accrue a diverse user community including urban planning and operations management researchers, as well as the economic and social policy community, in addition to public and private data government data organizations.

3. Urban Informatics

Overall, developments with urban Big Data have opened up several opportunities for urban analysis. Building on previous definitions (Foth et al., 2011; Bays and Callanan, 2012; Batty, 2013; Zheng, et al., 2014), we view Urban Informatics as the exploration and understanding of urban patterns and processes, and it involves analyzing, visualizing, understanding, and interpreting structured and unstructured urban Big Data for four primary objectives:

1) Dynamic resource management: developing strategies for managing scarce urban resources effectively and efficiently and often making decisions in real-time regarding competitive use of resources;
2) Knowledge discovery and understanding: discovering patterns in, and relationships among urban processes, and developing explanations for such trends;
3) Urban engagement and civic participation: developing practices, technologies and other processes needed for an informed citizenry and for their effective involvement in social and civic life of cities;
4) Urban planning and policy analysis: developing robust approaches for urban planning, service delivery, policy evaluation and reform, and also for the infrastructure and urban design decisions.

The overall framework used here, in terms of the objectives, research approach and applications, and their interdependencies, is shown in Figure 1.

**Figure 1: Relationships among Urban Informatics objectives, research approaches and applications**
3.1 Research Approaches in Urban Informatics

The analysis of urban systems is theoretically underpinned by myriad economic, social, behavioral, biological and physical principles, allowing the simulation of complex interactions, movements, transactions, trading, diffusion and other urban dynamics and diffusion patterns. While some urban models aim to improve long-range economic and infrastructural planning and program evaluation, others attempt to generate empirical understanding of urban dynamics and verification of theoretical urban concepts, and to provide input into shorter-term operations and management of urban sectors. However, Big Data has become closely associated with data-driven science and modeling, which is typically an empirical approach without the social, psychological, economic and regional planning theory which frame urban research. Data-driven modeling brings novel new methodological approaches particularly in using some of the highly unstructured and voluminous types of Big Data, and a bottom-up approach to understanding urban systems, particularly for improved dynamic resource management, knowledge discovery and citizen engagement.

The research approaches utilized in Urban Informatics are:

1) **Urban modeling and analysis with Big Data:** The use of Big Data within existing urban modeling and simulation frameworks, and in practical empirical approaches grounded in theoretical urban research paradigms, by: (a) reconfiguring/restructuring emerging Big Data through specialized data preparation techniques so that it meets the input requirements of existing urban modeling approaches; or (b) retrofitting or repurposing existing methods through the integration of data-driven approaches (e.g., machine learning, data mining) in the overall analysis scheme, so that urban models are able to use emerging forms of data.

2) **Data-driven models towards “bottom-up” sensing of the city:** Empirical data-driven methods that are derived primarily from the data science and statistical learning communities which focus on retrieval and extraction of information from unstructured or very voluminous streams of data that are not easily accessible to non-specialists, and their pattern detection, knowledge discovery, empirical explanation and hypothesis generation regarding urban phenomena, events and trends.

3.2 Urban Informatics Applications with Big Data

We organize the discussion on Urban Informatics applications using Big Data through urban models and data-driven models in the following ways: (1) reconsidering classical urban problems with new forms of data, (2) use of Big Data for complex systems analysis, (3) applications to address complex urban challenges through empirical research, and (4) through methods to collaboratively sense the city. The applications in turn, help to fine-tune the objectives of Urban Informatics for more comprehensive to knowledge discovery, urban planning and operations.

3.2.1 Reconsidering Classical Urban Problems with Big Data

Classical approaches to urban systems analysis include mathematical models of human spatial interaction to measure flows of travelers and services between pairs of points in urban areas (Wilson, 1971; Erlander, 1980; Sen and Srivastava, 1995), models of urban development, and study of urban structure, and the interaction between transportation and land-use systems (Burgess, 1925; Alonso, 1960; Lowry, 1964; Fujita and Ogawa, 1982; Fujita, 1988). Other areas are transportation network dynamics and travel mode choice analysis (Beckman, McGuire and Winston, 1956; Sheffi, 1985; Ben Akiva and Lerman, 1985), models of housing dynamics and residential location theory (Ellis, 1967; Muth, 1969; Beckman, 1973; Richardson, 1977); and models of regional and local economies, labor markets and industry location and agglomeration (Marshall, 1920; Isard, 1956; Krugman, 1991; Fujita et al., 1999). These methods are typically equation-based and draw from operations research and
statistical techniques. These models and their numerous variants have contributed to a voluminous and diverse literature on which several decades of planning, policy and operational decisions have been based.

These approaches typically use data from traditional sources such as the census or surveys, and to a lesser degree aggregated administrative or sensor data. Using many other urban Big Data sources would require significant modifications to such models “to see around the corners”, perhaps new model development or integration with data science approaches, in addition to specialist curation and processing of the data itself. However, there are now an increasing number of examples where emerging forms of data have been used within the framework of these classical urban models. Recent examples are in the areas of travel demand models, e.g., the use of GPS data to estimate flows of travelers from travel origins to destinations traditionally achieved using census journey-to-work data (Zhang et al., 2010), and the use of detailed freeway and arterial street sensor data along with the synthetic LEHD and other data to measure job accessibility (Levinson et al., 2010). Other examples include studies of labor market dynamics using administrative data (e.g., Bijwaard et al., 2011), use of social media data to measure labor market flows and indexes of job loss, job search, and job posting (Antenucci et al., 2014) and the use of online housing searches to study housing market dynamics in terms of area definition, submarket geography and search pressure locations (Rae, 2014).

3.2 Complex Systems Analysis

Large-scale urban modeling practice also use complex systems approaches utilizing Agent-based Models (ABM) and myriad forms of specialized survey, administrative, synthetic and other data sources, to study outcomes that are emergent from individual agent action in interaction with other agents and the environment while also incorporating agent heterogeneity. Well-known early implementations of ABMs include Schelling’s segregation model (Shelling, 1971) and Conway’s Game of Life (Conway, 1970). ABMs have found widespread application in diverse areas of urban research. Examples include urban and environmental planning (e.g. Zellner et al., 2009; Zellner and Reeves, 2012), transportation (e.g. Tilahun and Levinson, 2013; Zellner et al., forthcoming), environmental studies (e.g. Evans and Hugh, 2004), large-scale agent based micro-simulation models such as ILUTE (Salvini and Miller, 2005), and integrated land, transportation and environment modeling system such as MATSim (Balmer et al., 2009), which provides agent-based mobility simulations. Related developments in computational network perspectives to study a variety of phenomenon have also entered modeling practice, including studies of community structure (Girvan and Newman, 2002) and susceptibility of power grids to failure (Kinney et al., 2005).

ABMs have recently used unstructured sources of data, one example of which is the use of GPS trajectories to obtain a better understanding of human mobility patterns within an ABM framework (Jia et al., 2012). Some researchers have also focused on the use of social network data (Kowald and Axhausen, 2015 gives examples for the case of transportation planning), while others have utilized social networks to examine peer effects, and processes to exchange opinions, preferences and to share experiences, as well as to see how individual’s participation in social networks lead to outcomes of interest (e.g., Christakis and Fowler, 2007, demonstrated the spread of obesity via social relationships in a social network while Tilahun et al., 2011 examined the role of social networks in location choice). The use of online social networks in ABMs has been an interesting development in this respect allowing the flexibility of ABMs to incorporate detailed representation and analysis of the effects of social networks that underlie complex decision problems. One example of this nascent literature is the use of Twitter data within an ABM framework to model diffusion of crisis information (Rand et al., 2015).
3.3 Empirical Urban Research

A vast body of empirical work embedded in the urban disciplines is among the most active consumers of urban data, for better understanding, hypothesis testing and inference regarding urban phenomenon. Among these, one vast research area with requirements for specialized data sources, models and tools is that of environmental sustainability and issues relating to clean air, non-renewable energy dependence and climate change. While difficult to generalize, a recent OECD report identified gaps in quantitative urban and regional modeling tools to systematically assess the impacts of urban systems on climate change and sustainability (OECD, 2011). Significant developments in sensor technology have led to smart commodities ranging from household appliances to smart buildings leading to cost-efficiencies and energy savings, for the design of Vehicle-to-Grid (V2G) systems (Kempton et al., 2005), personal carbon trading (Bottrill, 2006) and vehicular cap-and-trade systems (Lundquist, 2011), with data-analytic research around technology adoption, impacts on behaviors and consumption patterns and so on.

Urban models that detect disparities relating to social justice and distributional aspects of transportation, housing, land-use, environment and public health are other consumers of such data. These approaches provide an empirical understanding of the social inclusion and livability aspects of cities, and operational decisions and policy strategies needed to address disparities. This line of work has focused on social exclusion and connections to work and social services (Kain and Persky, 1969; Wilson, 1987; Krivo and Petersen, 1996; Thakuriah et al., 2013), issues of importance to an aging society (Federal Interagency Forum on Aging-Related Statistics, 2010), health and aging in place (Black, 2008; Thakuriah et al., 2011) and needs of persons with disabilities (Reinhardt et al., 2011). Administrative data has played a significant role in this type of research leading to knowledge discovery about urban processes as well as in evaluation of governmental actions such as welfare reform and post-recession austerity measures. Linked and longitudinal administrative data can support understanding of complex aspects of social justice and changes in urban outcomes over time. For example, Evan et al. (2010) highlighted the importance of using longitudinal administrative data to understand the long-term interplay of multiple events associated with substance abuse over time, while Bottoms et al. (2009) discuss the role that longitudinal police crime records can play in studying repeat victimization of crime.

New ICT-based solutions to track and monitor activities allow urban quality and well-being to be assessed at more fine-grained levels. Personalized data generated by assistive technologies and ambient assisted living situations (Abascal et al., 2008) and other ICT applications can be used to assess contributory factors to urban quality of life as well as to design solutions supporting urban wellness objectives for seniors and persons with disability (e.g., hybrid qualitative-GPS data enabled as described by Huang et al., 2012 to understand barriers to accessing food by midlife and older adults with mobility disability). Mobile health and awareness technologies (Consolvo et al., 2006) particularly those embedded within serious medical pervasive gaming environments (e.g., DiabetesCity—Collect Your Data, Knoll, 2008) and numerous mobile, wearable and other sensor-based physical health recommender systems, one example of which is Lin et al. (2011), open up possibilities for urban researchers to tap into a wealth of data to understand overall built environment and activity-based conditions fostering health and well-being.

3.4 Approaches to Collaboratively Sense the City

The discussion above shows that urban information generation and strategies to analyze the data increasingly involve ICT solutions and the active participation of users. Strategies such as focus groups, SWOT, Strategic Approach, Future Workshops and other approaches have been extensively used in the past as a part of urban participatory practice to generate ideas and even to generate solutions to problems. However, advances in ICT solutions have led to the emergence of new models of citizens input into problem solving, plan and design sourcing, voting on projects, and sharing of
ideas on projects. Examples range from civic hackers analyzing data from Open Data portals to generate ideas about fixing urban problems to using serious games and participatory simulations for the ideation process (Poplin, 2014; Zellner et al., 2012).

As noted earlier, citizens may also engage by generating content through human computation, or by performing tasks that are natural for humans but difficult for machines to automatically carry out (von Ahn et al., 2008). Human computation approaches provide structured way for citizens to engage in play, to provide input and to interact with, and learn about the urban environment. For example, citizens may be able to judge different proposed urban design, or they may be used to assess the quality of urban spaces where objective metrics from data derived through machine vision algorithms are not accurate. Celino et al. (2012) gives an example of this called UrbanMatch, a location-based Game with a Purpose (GAWP), which is aimed at exploiting the experience that players have of the urban environment to make judgments towards correctly linking points of interests in the city with most representative photos retrieved from the Internet. There are multiple variants of human computation including social annotations (where users tag or annotate photos or real-world objects), information extraction (e.g., where users are asked to recognize objects in photos), and others.

By “sensing” the city and its different behavioral and use patterns, data-driven models have stimulated research into a broad range of social issues relevant to understanding cities, including building participatory sensing systems for urban engagement, location-based social networks, active travel and health and wellness applications, and mobility and traffic analytics. Other objectives include dynamic resource management of urban assets and infrastructure, assisted living and social inclusion in mobility, and community and crisis informatics. For example, one of the major cited benefits of social media analysis has been the ability to instantaneously and organically sense sentiments, opinions and moods to an extent not previously possible, and ways in which these diffuse over space and time, thereby enabling the policy community to monitor public opinion, and predict social trends. A part of this trend is being stimulated by major governmental agencies which are increasingly realizing the power of social media in understanding where needs are, and how the public are reacting to major policy changes and political events and people’s political preferences (Golbeck and Hansen, 2013).

A data-driven focus is also being seen in learning analytics (e.g., Picciano, 2012), location-based social networks (Zheng and Xie, 2011), recommender systems based on collaborative filtering for travel information (Ludwig et al., 2009) and approaches to detect disruptions from social media (Sasaki et al., 2012). Presumably if these information streams are collected over time and linked to other socio-demographic data, it would be possible to examine variations in the outcomes currently measured by the socially generated data to capture urban dynamics to a greater degree.

Overall, Big Data is being increasingly utilized for a range of Urban Informatics research and applications. By using existing urban models with new forms of data, or through data-driven modeling, urban processes and behaviors can be studied in a timely manner and contextual peculiarities of urban processes and local experiences can examined in greater detail. Yet significant challenges arise in their use, which are addressed next.

4. Challenges in Using Big Data for Urban Informatics

The challenges associated with the use of Big Data for Urban Informatics are: (1) technological, (2) methodological, (3) theoretical and epistemological, and (4) due to political economy that arise from accessing and using the data. These challenges are given in Table 2 along with the characteristics of the challenges and examples of the complexities involved with different types of Big Data.
### Table 2: Challenges in using Big Data for Urban Informatics and illustrative topics

<table>
<thead>
<tr>
<th>Challenges by type of data</th>
<th>Characteristics</th>
<th>Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Technological</strong></td>
<td>Urban information management challenges:</td>
<td>Information management challenges likely to very high with real-time, high-volume sensor and UGC data which require specific IT infrastructure development and information management solutions</td>
</tr>
<tr>
<td></td>
<td>1) Information generation and capture</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2) Management</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) Processing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4) Archiving, curation and storage</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5) Dissemination and discovery</td>
<td></td>
</tr>
<tr>
<td><strong>Methodological</strong></td>
<td>1) Data Preparation Challenges</td>
<td>Data preparation challenges likely to be very high with unstructured or semi-structured sensor, UCG and arts and humanities data, and data from real-time private-sector and administrative transactional systems</td>
</tr>
<tr>
<td></td>
<td>a) Information retrieval and extraction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b) Data linkage/information integration</td>
<td></td>
</tr>
<tr>
<td></td>
<td>c) Data cleaning, anonymization and quality assessment</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2) Urban Analysis Challenges</td>
<td>All types of observational Big Data pose significant methodological challenges in deriving generalizable knowledge requiring specialist knowledge to assess and address measurement issues and error structures</td>
</tr>
<tr>
<td></td>
<td>a) Developing methods for data-rich urban modeling and data-driven modeling</td>
<td></td>
</tr>
<tr>
<td></td>
<td>b) Ascertaining uncertainty, biases and error propagation</td>
<td></td>
</tr>
<tr>
<td><strong>Theoretical and epistemological</strong></td>
<td>1) Understanding metrics, definitions, concepts and changing ideologies and methods to understanding “urban”</td>
<td>All types of observational Big Data pose limitations in deriving theoretical insights and in hypothesis generation without adequate cross-fertilization of knowledge between the data sciences and the urban disciplines, but the challenges are greater with certain forms of UGC and sensor data which yield high-value descriptions but are less amenable to explanations and explorations of causality</td>
</tr>
<tr>
<td></td>
<td>2) Determining validity of approaches and limits to knowledge</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) Deriving visions of future cities and the links to sustainability and social justice</td>
<td></td>
</tr>
<tr>
<td><strong>Political economy</strong></td>
<td>1) Data entrepreneurship, innovation networks and power structures</td>
<td>Data confidentiality and power structures pose significant challenges to use of administrative data in open government and program evaluation, while access to private sector transactions data, and privately-controlled sensor and UGC are potentially susceptible to changing innovation and profitability motivations; challenges to ethics and responsible innovation are significantly high for certain sensor-based (e.g., IOT) applications</td>
</tr>
<tr>
<td></td>
<td>2) Value propositions and economic issues</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3) Data access, governance framework and provenance</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4) Data confidentiality, security and trust management</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5) Responsible innovation and emergent ethics</td>
<td></td>
</tr>
</tbody>
</table>

#### 4.1 Technological Challenges

Technological challenges arise due to the need to generate, capture, manage, process, disseminate and discover urban information. The challenges to managing large volumes of structured and unstructured information have been extensively documented elsewhere. Some of the major
information management challenges are those relating to building a data infrastructure, cloud stores and multi-cloud architectures, as well as resource discovery mechanisms, and language and execution environments. Other considerations include hardware, software, well-defined Application Programming Interfaces (API) needed to capture, integrate, organize, search and query and analyze the data. Of equal importance are scalability, fault-tolerance, and efficiency, and platforms for scalable execution. Various Big Data solutions have emerged in the market such as Hadoop, MapReduce and other solutions, some of which are open source.

One of the biggest challenges with using Big Data for Urban Informatics is not that the data are necessarily huge as in the case of financial, genomics, high-energy physics or other data (although this may change with the incoming deluge of the connected vehicle and the IoT world). Rather, it is that urban Big Data tends to be fragmented, messy and sometimes unstructured. Particularly for data linkage, when one goes beyond structured, rectangular databases to streaming data through APIs leading to text, image and other unstructured data formats, the diversity and fragmentation can pose significant problems.

Data privacy also becomes all-important with many sources of Big Data, whether they are administrative micro-data or user-generated image or GPS data, and is often a major roadblock to data acquisition for research, particularly for research that requires potentially personally identifiable data. There are many approaches to data privacy, and these range from technological encryption and anonymization solutions to design, access and rights management solutions. A vast range of Privacy Enhancing Technologies (PETs) (Beresford et al., 2003, Grusteser et al., 2003) are relevant to urban Big Data that focuses on anonymization of GPS data, images and so on. In the case of administrative micro-data, many approaches to ensure confidentiality are used, including de-identified data, simulated micro-data (called synthetic data) that is constructed to mimic some features of the actual data using micro-simulation methods (Beckman et al., 1996; Harland et al., 2012) and utilization of Trusted Third Party (TTP) mechanisms to minimize the risks of the disclosure of an individual’s identity or loss of the data (Gowans et al., 2012).

One major capability needed to progress from data-poor to data-rich urban models is that data should be archived over time, enabling storage of very high-resolution and longitudinal spatio-temporal data. The linkage to other socio-economic, land-use and other longitudinal data opens up additional avenues for in-depth exploration of changes in urban structure and dynamics. Although this was previously a challenge, decrease in storage costs and increase in linkage capacity has made this possible.

Another important determinant in data access is having access to high-quality resource discovery tools for urban researchers to find and understand data, ontologies for knowledge representation, and data governance framework that includes harmonization of standards, key terms and operational aspects. Given the vast and dispersed sources of urban Big Data, resource discovery mechanisms to explore and understand data are critical. This includes metadata or data about the data, containing basic to advanced information describing the data and the management rights to it, including archiving and preservation, in a consistent, standardized manner so that it is understandable and usable by others. Other issues are data lifecycle management (the strategic and operational principles underpinning long-term publication and archiving), access to necessary para-data, (i.e., data about the processes used to collect data), and social annotations (i.e., social bookmarking that allows users to annotate and share metadata about various information sources). These issues not only have significant technical requirements in terms of integrating urban data from different sources; they also have legal (e.g., licensing, terms of service, non-disclosure), ethical (e.g., regarding lack of informed consent in some cases, or use by secondary organizations which did not seek consent), and research culture implications (e.g., establishing a culture of reanalysis of evidence, reproduction and verification of
results, minimizing duplication of effort, and building on the work of others, as in Thanos et al., 2015).

The above technology issues are not likely to be directly relevant to urban researchers in many cases. However, methodological aspects of Big Data such as information retrieval, linkage and curation or the political economy of Big Data including data access, governance and privacy and trust management requirements may have direct implications for, and limit, urban researchers if appropriate technology solutions capable of handling these IT requirements cannot be found.

4.2 Methodological Challenges

We consider two types of methodological challenges: data preparation methods (such as cleaning, retrieving, linking, and other actions needed to prepare data for the end-user) and empirical urban analysis methods (data analytics for knowledge discovery and empirical applications). Sensor and co-created data require special processing and analysis methods to manage very large volumes of unstructured data, from which to retrieve and extract information. With traditional sources of urban data, the specific aspects of the workflow from data collection/generation to analysis are clearly demarcated among professionals from different backgrounds (e.g., data collection is typically done by census takers or surveyors who create a clean data file along with the necessary data documentation, which is then used by urban researchers for further analysis). In contrast to this model, in the case of certain forms of unstructured data (e.g., social media data such as Twitter), the analytics of the data (e.g., using machine learning for topic detection and classification algorithms) happens alongside with, or as a part of, information retrieval or the “gathering” of information from the raw data streams. Thus the “data gathering” and the “data analytics” aspects of the workflow are much more tightly coupled, requiring new skills to be learned by urban researchers wishing to use such data or to have close collaboration with data scientists who have this type of skills.

Observational Big Data involves having to address several methodological challenges for inference. Administrative data, for example, may pose challenges due to causality, endogeneity, and other issues that can bias inference. Socially generated data obtained from participatory sensing and crowdsourcing are likely to be non-representative in the sense that participants probably do not resemble random samples of the population. Those who are easiest to recruit may also have strong opinions about what the data should show and can provide biased information. Social media users are typically not representative of the overall population since they are more likely to be younger and more digitally savvy (Mislove et al., 2011), and they are also more likely to be concentrated in certain areas or generate differing content depending on where they live (Ghosh and Guha, 2013), although these patterns may change over time as the technology becomes more widely used.

In addition, technology changes rapidly and there would always be the issue of the first adopters with specific, non-representative demographics and use patterns. Aside from this, there is dominance by frequent issues and lack of data generation by passive users, and the proliferation of fake accounts which does not add real or true representation of moods, opinions and needs, and are sometime maliciously created to swell sentiments in one direction or the other. Other challenges include lack of independence or herding effects, which is the effect of prior crowd decisions on subsequent actions. Samples may need to be routinely re-weighted, again on the fly, with the weights depending on the purpose of the analysis. Recent work by Dekel and Shamir (2009), Raykar et al., (2010), and Wauthier and Jordan (2011) consider issues on sampling and sampling bias in crowd-sourcing or citizen science while others have considered sampling issues relating to social networks (Gjoka et al., 2010) and social media (Culotta, 2014). However, this work is in its infancy, and further developments are necessary in order to use highly unstructured forms of data for urban inference.

Using Big Data for Urban Informatics require methods for information retrieval, information extraction, GIS technologies, and multidisciplinary modeling and simulation methods from urban
research as well as the data sciences (e.g., machine learning and tools used to analyze text, image and other unstructured sources of data). Methods of visualization and approaches to understanding uncertainty, error propagation and biases in naturally occurring forms of data are essential in order to use and interpret Big Data for urban policy and planning.

4.3 Theoretical and Epistemological Challenges

The theoretical and epistemological challenges pertain to the potential for insights and hypothesis generation about urban dynamics and processes, as well as validity of the approaches used, and the limits to knowledge discovery about urban systems derived from a data focus. As noted earlier, Big Data for Urban Informatics has two distinct roots: quantitative urban research and data science. Although the walls surrounding what may be considered as “urban models and simulations” are pervious, these are typically analytical, simulation-based or empirical approaches that are derived from diverse conceptual approaches (e.g., queuing theory, multi-agent systems) and involve strong traditions of using specialist data to calibrate. These models support the understanding of urban structure, forecasting of urban resources, simulation of alternative investment scenarios, strategies for engagement of different communities, and evaluation of planning and policy, as well as efficient operations of transportation, environmental and other systems, using principles derived from theory. Such models are now using Big Data in varying degrees.

At the same time, exploratory data-driven research is largely devoid of theoretical considerations but is necessary to fully utilize emerging data sources to better discover and explore interesting aspects of various urban phenomena. Social data streams and the methods that are rapidly building around them to extract, analyze and interpret information are active research areas, as are analytics around data-driven geography that may be emerging in response to the wealth of geo-referenced data flowing from sensors and people in the environment (e.g., Miller and Goodchild, 2014). The timely discovery and continuous detection of interesting urban patterns possible with Big Data and the adoption of innovative data-driven urban management are an important step forward and serves useful operational purposes. The knowledge discovery aspects of data-driven models are important to attract the attention of citizens and decision-makers on urban problems and to stimulate new hypotheses about urban phenomena, which could potentially be rigorously tested using inferential urban models.

The limitation of the data-driven research stream is that there is less of a focus on the “why” or “how” of urban processes and on complex cause-and-effect type relationships. In general, data-driven methods have been the subject of interesting debates regarding the scope, limitations and possibility of such approaches to provide solutions to complex problems beyond pattern detection, associations, and correlations. The current focus on data-driven science and the advocacy for it have in some cases led to rather extreme proclamations to the effect that the data deluge means the “end of theory” and that it will render the scientific process of hypothesizing, modeling, testing, and determining causation obsolete (Anderson, 2008). Quantitative empirical research has always been a mainstay for many urban researchers but there is inevitably some conceptual underpinning or theoretical framing which drive such research. Long before Big Data and data science became options to derive knowledge, the well-known statistician, Tukey (1980), noted in an article titled “We Need Both Exploratory and Confirmatory” that “to try to replace one by the other is madness”, while also noting that “ideas come from previous exploration more often than from lightning strikes”.

A part of the debate is being stimulated by the fact that data-driven models have tended to focus on the use of emerging sources of sensor or socially co-created data, and is closely connected to the data science community. At the time of writing this paper, entering “GPS data” into the Association for Computing Machinery (ACM) Digital Library, a major computer science paper repository returns about 11,750 papers, while entering the same term in IEEE Xplore Digital Library, another such
source, returns another 6,727 papers; these numbers are in fact higher than the counts obtained when the first author was writing her book “Transportation and Information: Trends in Technology and Policy” (Thakuriah and Geers, 2013), indicating not just a voluminous literature on these topics in the data sciences but one that continues to grow very fast. Such data sources have become most closely associated with the term Big Data in the urban context, to the exclusion of administrative data, hybrids of social survey and sensing data, humanities repositories and other novel data sources, which play an important role in substantive, theoretically-informed urban inquiry, beyond detection, correlations and association.

Sensor and ICT-based UGC has also become closely associated with smart cities, or the use of ICT-based intelligence as a development strategy mostly championed and driven by large technology companies for efficient and cost-effective city management, service delivery and economic development in cities. There are numerous other definitions of smart cities, as noted by Hollands (2008). The smart cities movement has been noted to have several limitations, including having “a one-size fits all, top-down strategic approach to sustainability, citizen well-being and economic development” (Haque, 2012) and for being “largely ignorant of this (existing and new) science, of urbanism in general, and of how computers have been used to think about cities since their deployment in the mid-20th century” (Batty, 2013), a point also made by others such as Townsend (2013). It needs to be pointed out that the scope of smart cities has expanded over time to include optimal delivery of public services to citizens and on processes for citizen engagement and civic participation, as encapsulated by the idea of “future cities”.

Urban Big Data is also now being strongly associated with Open Data; Open Data is now being increasingly linked to smart cities, along with efforts to grow data entrepreneurship involving independent developers and civic hackers to stimulate innovations and promote social change. Nevertheless, at least in the European Union, the European Innovation Partnership (EIP) on Smart Cities and Communities has received some 370 commitments to fund and develop smart solutions in the areas of energy, ICT and transport. These commitments involve more than 3,000 partners from across Europe towards creating “a huge potential for making our cities more attractive, and create business opportunities” (European Commission, 2015). It is little wonder that the term Big Data for cities is being referred to in some circles almost exclusively in the context of smart cities, to the exclusion of urban research contributions, including a long-standing urban operations literature using sensor and at least some types of user-generated data.

Notwithstanding these tensions, some of the benefit of using sensor and socially generated forms of Big Data is in identifying contextual particularities and local experiences that are very often smoothed over by the systems-oriented view of quantitative urban research; the latter often emphasizes generalizability, sometimes masking elements of complex urban challenges. Such “local” focus lends the hope that observations of unique local features from data will stimulate interest in exploring previously unknown hypothesis of urban processes and that the unique local urban problems identified potentially lends itself to context-dependent urban policy and plan-making. The “new geographies of theorizing the urban” (Robinson, 2014, Roy, 2009) is oriented to skepticism regarding authoritative and universalizing claims to knowledge about urban experiences and is committed to giving attention to contextual particularities and local experiences within places (Brenner and Schmid, 2015). Although epistemological links between data-driven urban modeling and critical urban theory is virtually non-existent at the current time, and may never be explicitly articulated, novel sources of Big Data have the potential to allow the capture of data on social, behavioral and economic aspects of urban phenomena that have either not been previously measured or have been measured at resolutions that are too aggregated to be meaningful. However, such localized observations are far from being a substitute for qualitative social science research, as noted by Smith (2014), who advocates a continued need for ethnographic approaches and qualitative methods and cautions against the continued separation of method from methodology and discipline.
Further, causality assessment is challenging with many forms of Big Data and it does not lend itself easily to the derivation of counterfactuals and to forming an etiologic basis for complex urban processes. Instead of moving from using urban models to a completely different data-driven era, as noted earlier, the focus may be to shift to using administrative, sensing or socially generated urban Big Data as input into estimating and testing traditional models. Urban Big Data analysis would also benefit from being linked to behavioral models needed to build alternative scenarios to understand the effects of unobserved assumptions and factors, or to derive explanations for parts of the urban environment not measured by data. The linked data hybrids suggested previously potentially offers a way to address these limitations.

4.4 Challenges due to the Political Economy of Big Data

The political economy of Big Data arises due to the agendas and actions of the institutions, stakeholders and processes involved with the data. Many of the challenges facing urban researchers in using Big Data stem from complexities with data access, data confidentiality and security, and responsible innovation and emergent ethics. Access and use conditions are in turn affected by new types of data entrepreneurship and innovation networks, which makes access easier in some cases through advocacy for Open Data or makes it more difficult through conditions imposed as a result of commercialization and are generally underpinned by power structures and value propositions arising from Big Data.

The economic, legal and procedural issues that relate to data access and governance are non-trivial and despite the current rhetoric around the open data movement, vast collections of data that are useful for urban analysis are locked away in a mix of legacy and siloed systems owned and operated by individual agencies and private organizations, with their own internal data systems, metadata, semantics and so on. Retrieving information from social media and other online content databases, and the analytics of the resulting retroactive UGC either in real-time or from historical archives have mushroomed into a significant specialized data industry, but the data availability itself is dictated by the terms of service agreements required by the private companies which own the system or which provide access, giving rise to a new political economy of Big Data. User access is provided in some cases using an API, but often there are limits on how much data can be accessed at any one time by the researcher and the linkage of a company’s data to other data. Others may mandate user access under highly confidential and secure access conditions requiring users to navigate a complex legal landscape of data confidentiality, and special end-user licensing and terms of service and non-disclosure agreements. Data users may also be subject to potentially changing company policy regarding data access and use. There are also specific restrictions on use including data storage in some cases, requiring analytics in real-time.

Undoubtedly, a part of the difficulty in access stems from data confidentiality and the need to manage trust with citizens, clients and the like. Privacy, trust and security are concepts that are essential to societal interactions. Privacy is a fundamental human right and strategies to address privacy involve privacy-enhancing technology, the legal framework for data protection, as well as consumer awareness of the privacy implications of their activities (Thakuriah and Geers, 2013), especially as users leave a digital exhaust with their everyday activities. However, privacy is also not a static, immutable constant. People are likely to trade off some privacy protection in return for utility gained from information, benefits received, or risks minimized (Cottrill and Thakuriah, 2015). Aside from technological solutions to maintain privacy, a process of user engagement is necessary to raise consumer awareness, in addition to having the legal and ethics processes in place in order to be able to offer reassurance about confidential use of data. Further, many private data owners may not release data due to being able to reserve competitive advantage through data analytics. However, lack of knowledge about the fact-moving legal landscape with regards to data confidentiality, copyright
violations and other unintended consequences of releasing data are central elements of the political economy of Big Data.

The social arguments for and against Big Data, connected systems and IoT are similar to other technology waves that have been previously witnessed, and these considerations also generate multiple avenues for research. Increasingly pervasive sensing and connectivity associated with IoT, and the emphasis on large-scale highly coupled systems that favor removing human input and intervention has been seen to increase exposure to hacking and major system crashes (BCS, 2013). Aside from security, the risks for privacy are greatly enhanced as the digital trail left by human activities may be masked under layers of connected systems. Even those systems that explicitly utilize privacy by design are potentially susceptible to various vulnerabilities and unanticipated consequences since the technological landscape is changing very rapidly and the full implications cannot be thought through in their entirety. This has prompted the idea of “responsible innovation”, which “seeks to promote creativity and opportunities for science and innovation that are socially desirable and undertaken in the public interest” and which makes clear that “innovation can raise questions and dilemmas, is often ambiguous in terms of purposes and motivations and unpredictable in terms of impacts, beneficial or otherwise. Responsible Innovation creates spaces and processes to explore these aspects of innovation in an open, inclusive and timely way” (Engineering and Physical Sciences Research Council, n.d.).

Against this backdrop of complex data protection and governance challenges, and the lure of a mix of objectives such as creating value and generating profit as well as public good, a significant mix of private, public, non-profit and informal infomediaries, ranging from very large organizations to independent developers that are leveraging urban Big Data have emerged. Using a mixed-methods approach, Thakuriah et al. (2015) identified four major groups of organizations within this dynamic and diverse sector: general-purpose ICT companies, urban information service providers, open and civic data infomediaries, and independent and open source developer infomediaries. The political economy implication of these developments is that publicly available data may become private as value is added to such data, and the publicly-funded data infrastructure, due to its complexity and technical demands, are increasingly managed by private companies that in turn, potentially restricts access and use.

5. Conclusions

In this paper, we discussed the major sources of urban Big Data and their benefits and shortcomings, and ways in which they are enriching Urban Informatics research. The use of Big Data in urban research is not a distinct phase of a technology but rather a continuous process of seeking novel sources of information to address concerns emerging from high cost or design or operational limitations. Although Big Data has often been used quite narrowly to include sensor or socially generated data, there are many other forms that are meaningful to different types of urban researchers and user communities, and we include administrative data and other data sources to capture these lines of scholarship. But even more importantly, it is necessary to bring together (through data linkage or otherwise), data that have existed in fragmented ways in different domains, for a holistic approach to urban analysis.

We note that both theory-driven as well as data-driven approaches are important for Urban Informatics but that retrofitting urban models to reflect developments in a data-rich world is a major requirement for comprehensive understanding of urban processes. Urban Informatics in our view is the study of urban patterns using novel sources of urban Big Data that is undertaken from both a theory-driven empirical perspective as well as a data-driven perspective for dynamic resource management, knowledge discovery and understanding, urban engagement and civic participation, and urban planning and policy. The research approaches utilized to progress these objectives are a mix of enriched urban models underpinned by theoretical principles and retrofitted to accommodate emerging forms of data, or data-driven modeling that are largely theory-agnostic and emerge bottom-
up from the data. The resulting Urban Informatics research applications have focused on revisiting classical urban problems using urban modeling frameworks but with new forms of data; evaluation of behavioral and structural interactions within enriched complex systems approach; empirical research on sustainable, socially-just and engaged cities; and applications to engage and collaboratively sense cities.

The use of Big Data pose considerable challenge for Urban Informatics research. This includes technology-related challenges putting requirements for special information management approaches, methodological challenges to retrieve, curate and draw knowledge from the data; theoretical or epistemological challenges to frame modes of inquiry to derive knowledge and understand the limits of Urban Informatics research; and finally, an issue that is likely to play an increasingly critical role for urban research – the emerging political economy of urban Big Data, arising from complexities associated with data governance and ownership, privacy and information security, and new modes of data entrepreneurship and power structures emerging from the economic and political value of data.

From the perspective of urban analysts, the use of sensor data, socially generated data, and certain forms of arts and humanities and private sector data may pose significant technical and methodological challenges. With other sources such as administrative micro-data, the data access challenges and issues relating to political economy and data confidentiality might be non-trivial. Issues such as sustainability of the data infrastructure, dealing with data quality, and having access to the skills and knowledge to make inferences, apply to all forms of naturally occurring data.

While many types of urban Big Data such as administrative data and specific sensor systems have been used for a long time, there are many novelties as well, such as new, connected sensor systems, and socially generated or hybrid, linked data that result in data in new formats or structure. There is a need for a wide variety of skills due to the tight coupling of preparing unstructured data and data analysis, but also due to the wide variety of technological, methodological and political economy issues involved. Additionally, data and analytics are only one part of the data-focused approach to urban operations, planning and policy-making; having the mechanisms to interpret the results and to highlight the value derived, is critical for adoption of data-driven strategies by decision-making, and for its eventual impact on society.

It is therefore an opportune time for an interdisciplinary research community to have a discussion on the range of issues relating to the objectives of Urban Informatics, the research approaches used, the research applications that are emerging, and finally, the many challenges involved in using Big Data for Urban Informatics.

**References**


24


Developing An Interactive Mobile Volunteered Geographic Information Platform to Integrate Environmental Big Data and Citizen Science In Urban Management

Yanfu Zhou¹, Zhenghong Tang¹*, Hongfeng Yu², Yue Gu¹, Tiantian Liu¹

¹ Community and Regional Planning Program, University of Nebraska-Lincoln
² Department of Computer Science and Engineering, University of Nebraska-Lincoln
(*corresponding author: Phone: (402)472-9281; Email: ztang2@unl.edu)

Abstract: A significant technical gap exists between the large amount of complex scientific environmental big data and the limited accessibility to these datasets. Mobile platforms are increasingly becoming important channels through which citizens can receive and report information. Mobile devices can be used to report Volunteered Geographic Information, which can be useful data in environmental management. This paper evaluates the strengths, weaknesses, opportunities, and threats for six cases: “Field Photo,” “CoCoRaHS,” “OakMapper,” “What’s Invasive!”, “Leafsnap,” and “U.S. Green Infrastructure Reporter.” The results indicate that active, loyal and committed users are key to ensuring the success of citizen science projects. Online and off-line activities should be integrated to promote the effectiveness of public engagement in environmental management. It is an urgent need to transfer complex environmental big data to citizens’ daily mobile devices which will then allow them to participate in urban environmental management. A technology framework is provided to improve existing mobile-based environmental engagement initiatives.

Keywords: environmental big data, citizen science, urban environmental management, mobile, Volunteered Geographic Information

Introduction

Professional environmental datasets such as the Soil Survey Geographic database (SSURGO), Flood Insurance Rate Map (FIRM), National Wetland Inventory (NWI), and the water quality dataset (STORET) provide centralized environmental information on a nationwide basis. Although more accurate, detailed, geo-referenced, real-time information is being collected on a daily basis, these datasets are increasingly becoming much larger and more technical. At the same time, technical barriers still exist for the general public to be able to access nearby environmental information. Specific software (e.g., ArcGIS) and certain technical skills are needed to read these datasets. In fact, professional educators, researchers, planners, or urban managers often have difficulty in accessing some professional datasets (e.g., SSURGO data, atrazine pollution datasets) in the field, which may require using additional expensive equipment (e.g., a hand-held GPS unit). A limited number of user-friendly systems are available on mobile platforms in spite of the fact that mobile devices are rapidly becoming a predominant information channel, particularly for younger generations including college students. Most existing big environmental datasets are only accessible through websites or even hard copies. The major information channel used by our younger generations is dramatically shifting to GPS-enabled mobile devices, this critical time also represents a great opportunity to transfer big environmental datasets to mobile platforms. A mobile information platform can not only improve public awareness
of environmental conditions in a community, but it can also improve the efficiency and effectiveness of how environmental data are used in urban management.

More importantly, most of these environmental datasets are still read-only in nature, whereby one can only view the data. There is no real-time reporting functionality to enable citizen science. In addition, many of the professional environmental datasets (e.g., NWI, FIRM, and SSURGO) have inaccurate or out-of-date information. Today, citizen science is recognized as a critical resource in verifying and updating environmental data. However, citizens have inadequate tools or channels with which to share their observed information with stakeholders such as educators, researchers, and managers. From a crowdsourcing perspective, citizen scientists have not been fully empowered to participate in the traditional urban environmental management framework. However, mobile mapping tools, geo-tagged social networks, and GPS-enabled mobile devices provide robust tools for collecting real-time environmental data. With these rapidly developing technologies, citizen science is well positioned to make even more important contributions to environmental management, including air quality monitoring, water quality monitoring, and biodiversity conservation.

This paper uses a case study methodology to review pioneering crowdsourcing environmental information platforms. A technical framework with learned experiences and lessons from two federally-funded projects and the case studies are provided to further integrate big environmental data with citizen science in urban environmental management. The SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis methodology is used to qualitatively evaluate user, information, and site use of these cases. A technical framework is provided to guide the future development of citizen science projects.

**VGI in Environmental Management**

Volunteered Geographic Information (VGI) is a special case of the more general web phenomenon of user-generated content. The term VGI was first coined by Michael F. Goodchild in 2007 (Goodchild, 2007). Volunteered behavior, whether performed by an individual or an anonymous group, has existed in human history for a long time, and played a significant role when humans first sought to control nature. Although it remains unclear why people volunteer themselves, the concept of VGI and citizen science has been applied to many research fields and business areas, and has resulted in countless positive effects in society (Goodchild and Glennon, 2010; Jordan et al., 2012; Devictor et al., 2010; Dragicevic and Balram, 2006; Elwood, 2008; Elwood, 2010; McCall, 2003; Werts, 2012).

For example, in 1854, Dr. John Snow, a volunteer doctor, discovered the reason behind the fact that cholera had caused massive deaths in London, when he showed that a central water source was responsible for the outbreak of cholera on his map (Johnson, 2006). However, because there were no modern technologies available at that time, it took a long time to spread snow's idea to the general public. With modern technologies, it is possible for an individual to distribute his/her wisdom or idea over the world quickly through social media such as Facebook, Wikipedia, Twitter, and Flickr.

Overall, there have been many successful volunteered information cases ranging from the 19th century to the Internet era. We note that the spatial property enabled by today’s Web 2.0 technology can be regarded as a special enhancement for volunteered information, which is under the umbrella of VGI. VGI methods are a type of crowdsourcing, and the discussion between VGI and crowdsourcing cannot be divided. The practice of
VGI will most likely improve the theory and methodology of crowdsourcing. This theory could also be a guide for using VGI methods. In addition, the crowdsourcing data generated by VGI users can provide invaluable results for decision-making.

The integration of VGI with environmental monitoring and management allows users, whether they are citizens or scientists, to create new databases or maintain and update existing databases. However, existing environmental databases are typically not user-friendly and are difficult for inexperienced users to access. User interfaces of existing databases are usually outdated because most database websites were built in the early stage of the Internet, and their web interfaces are not compatible with advanced devices such as smart phones or tablets. Also, some existing databases are massive; if users want to access the data, they must follow the instructions on the websites, download the data, and then open the databases with professional tools.

Authoritative information and tools, such as remote sensing data from NASA (NASA, National Aeronautics and Space Administration) or NWI data from USWFS (NWI, a product of the U.S. Fish & Wildlife Service), are accurate, of high resolution, and reliable. However, they are expensive, time-consuming, inaccessible, and geographically limited. In contrast, using VGI to aid environmental monitoring and management does not require expensive updating and maintenance for high-resolution remote sensing data. This is because users, instead of agencies and corporations, update the maps. VGI also does not incur the costs of building large databases, because once a VGI system is built, it functions as a data-driven website (data-driven website, one of the types of dynamic web pages), much like Zillow.com (Zillow.com, a product of Zillow, Inc., an online real estate data-driven website). In addition, VGI can also bridge the gap between citizens, scientists, and governments. The application of VGI in environmental monitoring also enables neogeography, which emphasizes the importance of participants in knowledge production, reducing existing gaps between the public, researchers, and policymakers (Peluso, 1995; Bailey et al., 2006; Mason and Dragicevic, 2006; Parker, 2006; Walker et al., 2007). Applying VGI in environmental monitoring enables access to potential public knowledge (Connors et al., 2012). It is only limited by a user’s spatial location, and therefore it is more flexible than traditional methods in certain cases such as flood rescue response and water quality management.

While some users may just want to browse rather than modify data, the learning process required by accessing existing databases may reduce a user’s motivation and become very time-consuming. For example, if a user wants to check whether his or her backyard is in a 100-year flood zone area when standing outside his or her house, the user doesn’t need to do extensive research on FEMA’s website and open data from ESRI’s ArcMap program (ESRI, Environmental Systems Research Institute). Instead, the user can simply access FEMA data on a smartphone anytime, anywhere. The enabled Web 2.0 technologies and VGI methods can resolve these previously-mentioned issues, incorporate websites to be compatible with smart devices, and transfer databases to crowdsourcing clients, which would significantly benefit environmental monitoring and management.

**Criteria for Case Selection**

In order to ensure the objectivity of the case study methodology, this paper adopts six criteria for case selection: information platform, addressed issue, data collection method, data presentation, founder, and coverage. The information platform of the target cases
should have interactive mobile-accessible platforms to allow citizens to view the spatial
information and engage their participation through mobile devices. Selected cases should
be environmentally-related topics. Data collection should rely on the public's submissions
as primary data sources. The data presentation should have geospatial maps. The funders
should involve universities as developers. The coverage indicates each case should address
a different topic to avoid repeatability and ensure the diversity of selected cases. Based on
the above criteria, six cases were selected: “Field Photo”, “CoCoRaHS,” “OakMapper,”
“What’s Invasive!”, “Leafsnap,” and “U.S. Green Infrastructure Reporter.”

Case Studies

“Field Photo” was developed by the Earth Observation Modeling Facility at the
University of Oklahoma. It has a mobile system to enable citizens to share their field photos,
show footprints of travel, support monitoring of earth conditions, and verify satellite image
data. Field Photo can document citizen observations of landscape conditions (e.g., land use
types, natural disasters, and wildlife). Citizen-reported photos are included in the Geo-
Referenced Field Photo Library which is an open-sourcing data archive. Researchers, land
managers, and citizens can share, visualize, edit, query, and download the field photos.
More importantly, these datasets provide crowdsourcing geospatial datasets for research
on land use and land cover changes, the impacts of extreme weather events, and
environmental conservation. Registered users have more accessibility to the photo library
than guest users do. Both the iOS and Android versions of “Field Photo” applications have
been available since 2014 for citizen download.

“CoCoRaHS” represents the Community Collaborative Rain, Hail and Snow
Network that was set up and launched by three high school students with local funding.
CoCoRaHS is a non-profit, community-based network of volunteers who collaboratively
measure and map precipitation (rain, hail and snow) (Cifelli et al., 2005). Beginning with
several dozens of enthusiastic volunteers in 1998, the number of participants has increased
every year. Besides active volunteers, there are some people who have participate in this
program for a few weeks but have not remained active over the long-term. In 2000,
CoCoRaHS received funding from the NSF’s Geoscience Education program and was
operated by the Colorado Climate Center at Colorado State University. Based on real-time
statistical data, around 8,000 daily precipitation reports were received in 2013 across the
United States and Canada. Mobile applications of CoCoRaHS Observer for iOS and
Andriod systems were provided by Steve Woodruff and Appcay Software (not CoCoRaHS)
to allow registered volunteers to submit their daily precipitation reports via their mobile
devices. Registration is required to use these mobile applications.

“OakMapper” was developed by the University of California-Berkeley in 2001
(Kelly et al., 2003). Sudden oak death is a serious problem in California and Oregon forests.
Because there are so many people who walk or hike in these forests, “OakMapper”
extended its original site, which further allowed communities to monitor sudden oak death.
“OakMapper” has successfully explored the potential synergy of citizen science and expert
science efforts for environmental monitoring in order to provide timely detection of large-
scale phenomena (Connors et al., 2012). By 2014, it had collected 3,246 reports, most of
which came from California. However, “OakMapper” is not a full real-time reporting
system. Submitted data can only be displayed if it contains a specific address. As of 2014,
OakMapper only had an iOS-based mobile application. Users can view the data, but they need to register in order to submit volunteered reports.

“What’s Invasive!” is a project that attempts to get volunteered citizens to locate invasive species anywhere in the United States by making geo-tagged observations and taking photos that provide alerts of habitat-destroying invasive plants and animals. This project is hosted and supported by the Center for Embedded Networked Sensing (CENS) at the University of California, Los Angeles, the Santa Monica Mountains National Recreation Area, and EDDMapS (Invasive Species Mapping Made Easy), a web-based mapping system for documenting invasive species distribution developed by the University of Georgia's Center for invasive Species and Ecosystem Health. Any user who registers as a project participant must provide an accurate email address. Users can self-identify as a beginner, having some identification training, or an expert according to their knowledge levels. This project only tracks statistical data such as the frequency with which users log in for research use. No personal background data on users are collected. Both the iOS and Android versions of mobile applications have been available since 2013 for citizen download. “Leafsnap” is a pioneer in a series of electronic field guides being developed by Columbia University, the University of Maryland, and the Smithsonian Institution. It uses visual recognition software to help identify tree species from photographs of their leaves. Leafsnap provides high-resolution images of leaves, flowers, fruit, petiole, seed, and bark in locations that span the entire continental United States. Leafsnap allows users to share images, species identifications, and geo-coded stamps of species locations and map and monitor the ebb and flow of flora nationwide. Both the iOS and Android versions of mobile applications are available for citizen use.

“U.S. Green Infrastructure Reporter” was developed in the Volunteered Geographic Information Lab at the University of Nebraska-Lincoln in 2012. Its main purpose is to allow stakeholders and citizens to report green infrastructure sites and activities through their mobile devices. This mobile information system has a GPS-synchronous real-time reporting function with its own geospatial database that can be used for analysis. It provides both iOS and Android versions of mobile applications. More than 6,700 reports were collected across the United States by 2013.

SWOT Analysis

This study adopts a qualitative analysis method to analyze the selected cases. SWOT (Strengths, Weaknesses, Opportunities, and Threats analysis) is a structured method used to evaluate these cases. The strengths indicate the advanced characteristics of one case over other cases. The weaknesses indicate the disadvantages of a case in this usage. Opportunities mean the elements in an environment that can be exploited to their advantage. Threats indicate the elements in an environment that could cause trouble or uncertainty for development. SWOT analysis results can provide an overview of these cases in terms of their existing problems, barriers, and future directions.

Results
**Strengths:** The mobile-based information platform has more unique strengths than other web-based platforms. All the cases were built on mobile platforms. Citizens can use their portable mobile devices to participate in these projects and do not need to rely on special devices. The project can be incorporated into their daily activities. A user-friendly participatory tool is helpful in solving large-scale and long-term environmental problems. The significant strengths of these selected cases clearly address the specific tasks. All of these cases address very well-defined topics that are easily understood by citizens. When citizens find topics that fit their interests and have the opportunity to contribute to a problem-solving process, they tend to convert their interests into real action in order to participate in these projects. Compared with web-based technologies, mobile information platforms can transcend the limitations of time and space and thus attract grassroots participation. Citizens can use GPS-enabled mobile devices to report geo-referenced information anytime, anywhere. Compared to traditional methods, VGI has many advantages for environmental management including reduced costs and adaptability. The crowdsourcing data generated by volunteers can be an alternative data source for environmental management; there is only a slight difference in quality between the data collected by experts and the data collected by non-experts. Additionally, VGI can be used as an effective means to promote public awareness and engagement, compensating for the weaknesses of traditional methods.

Although the VGI method will not replace traditional methods, it will augment traditional methods for environmental management. VGI can also be considered as a new way to rapidly collect information from third parties, nonprofit organizations, and municipal departments. The databases created by VGI users can provide multiple benefits for environmental management in future studies such as flood mitigation research. When designing an application for a VGI method, planners also need to think about people's urgent and daily needs, which could be very helpful for planners to reach the public and improve citizen engagement. In addition, using VGI methods could accelerate the planning process and save costs, since residents can contribute to part of the planning process instead of the government. The strengths are summarized below:

- With a user-friendly, easily-accessible, convenient platform
- With a clearly-defined specific task
- Attract a large amount of participants
- With GPS-synchronous information
- Attraction for a large amount of users
- Grassroots reports and information
- Not limited to submission time and place

**Weaknesses:** The weaknesses of these projects include limited availability of mobile devices, age gaps, data quality issues, and data verification issues. Many citizens still cannot afford GPS-enabled mobile devices with expensive data plans. The relatively elderly population has less knowledge and/or interest in using mobile devices. The quality of citizen-reported data is a fundamental challenge for any citizen science project. Data verification needs a significant amount of time and resources. Experiences with these projects also bear out that the number of users does not equal the number of active users. Most of the data may only be submitted by a small number of active users. The total number of users only provides an overview of project coverage, but the number of registered users
is a better indicator of those who contribute to citizen science projects. In addition, a high level of participation can bring a large amount of citizen-submitted data, but this does not mean that citizens provide a large amount of useful data. Data quality comes from users’ understanding, judgment and operation experiences. The quality verification depends on an expert’s knowledge and experience. An automated filter may miss some unusual observations, but these observations might be real. A manual verification procedure, on the other hand, can reduce the probability of error, but needs more time and money to implement. Compared to the authoritative data from agencies or corporations, the quality of data collected from VGI systems is typically a concern. However, it has been proven that there is no significant difference between data collected by scientists and volunteers and, in fact, volunteers can make valuable contributions to the data collection process.

The VGI method has similar problems to other crowdsourcing practices. Although there is no significant difference between the data collected by scientists and the data collected by citizens, the VGI method still has occasional data quality problems, especially when volunteers are not motivated to contribute their knowledge to the VGI system. Some volunteers may not be willing to finish all the inventory questions when the results of the VGI methods cannot be seen immediately. A major barrier to environmental protection measures is that it usually takes decades for people to see the results of their environmental stewardship. In addition to data quality issues, crowdsourcing also has the following problem: There is no time constraint for the data collection process with the VGI method. When using the VGI method, it is hard for planners to define or identify what citizens’ motivations and interests are, and whether their motivation and interests will have a positive or a negative impact to the planning process. How to prohibit intellectual property leakage for sensitive topics, such as using the VGI method to collect information on housing prices, is another major concern. There are currently no standards for designing a VGI system, so there is not much control over the development or the ultimate product. When using the VGI method, planners also need to think about what things can be crowd sourced and what cannot, such as content that has copyright issues. The weaknesses are summarized as below:

- Limited availability of mobile devices
- Age gaps
- Data quality issues
- Data verification issues
- Lack of registered users
Opportunity: Mobile-based projects provide a new avenue to participate in environmental management. Mobile devices are currently very popular as a part of people’s daily lives. Mobile information platforms are increasingly becoming the predominant information channel for people to receive and report information. Many of these projects, such as “Field Photo” and “Leafsnap,” have successfully combined online activities and off-line engagement activities to attract new participants and retain older users. The case study also found that face-to-face communication is a valuable portion of mobile engagement. The integration of nonverbal communication and personal touches can improve the effectiveness of mobile engagement. In addition, volunteered coordinators are helpful in promoting volunteer activities. Social media and social networks are promising channels through which to attract people in the virtual community. These mobile VGI projects are still not empowered by social media and social networks. Planners also need to think about how to appreciate and recognize contributions from the public and how to preserve existing VGI users. In general, applying VGI methods to planning means not only building a VGI system. It is both a planning process and a crowdsourcing practice. Crowdsourcing can be considered as a new and upcoming planning issue in the future, especially when e-government practices are expected to become more popular in the future. The opportunities are summarized as below items:

- More popular for mobile devices
- Mobile devices as the first information channel
- Combination of online and off-line activities
- Empowerment of traditional communication techniques
- Integration with social media
- Collaboration with other organizations

Threats: The first threat is limited motivation from citizens (Agostinho and Paço, 2012). Projects without adequate motivation are the greatest threat to long-term success. Most citizens have no interest in scientific research projects that cannot bring any direct economic benefits to them. A successful engagement should build a shared vision and devote responsibility and commitment in specific tasks. Incentives can be useful to attract long-term, active contributors. A two-way communication platform cannot be fully engaged due to lack of timely feedback from the organizers. The reported data can only be directly used by the project organizers for analysis. Current databases do not have efficient approaches to manage citizen science data such as photos and videos. The analysis of the collected data is still very superficial and does not have any in-depth data mining. The collected data types mainly only include text and photos and do not include more scientific data. These projects still lack strategic management and sustainability. Most of these projects depend on specific funding to support their daily operations. Long-term management strategies with alternative funding resources are necessary to reduce the threats and ensure effective implementation.

Although technology has improved greatly on mobile devices, there are still two technology problems that have an impact on adopting VGI methods: signal coverage range and signal quality for mobile devices, and battery life. Although most mobile devices can get good signals in urban areas, in rural areas mobile devices can disconnect suddenly due to inadequate signal range or poor signal quality, which can result in missing or erroneous data. In addition, information service providers have different signal coverage. Even in the
same city, the signal quality and strength varies greatly and VGI users may lose patience if their phone service isn’t available in a particular area. Signal quality and strength not only has an impact on the adoption of VGI methods, it also has an impact on the volunteers themselves. Adopting VGI methods also requires users to have better mobile devices, although VGI developers and planners will try to cover as many popular devices as they can and reduce memory requirements for mobile applicants.

There are still some hardware limits that cannot be solved, however, such as battery life. If the VGI users are a group of professionals, they may use VGI apps on their mobile devices to collect data for an entire 24 hours, and the screen will be constantly refreshing, reducing the battery life. Another problem with adopting VGI methods is covering the entire mobile phone market. Currently Google and Apple take the majority of the smartphone market. It is in the VGI developers’ best interest to have their apps deployed on both Google and Apple phones. This is not a simple task because Google and Apple have different policies for reviewing a developer’s app. Google is open source and allows more freedom to access phone abilities, whereas Apple is closed and is sensitive to accessing phone abilities such as sensors. Trying to keep VGI apps working both on Google and Apple phones is a challenge for developers and planners. In general, Google and Apple both provide a good development space for promoting VGI methods, and each has its own advantages and disadvantages.

The invasion of hackers is still a threat, and the same concerns extend to the mobile era. From a technological viewpoint, it is true that some smartphone systems have safety issues, and hacking software targets to smart phones is common. Utilizing VGI methods can also become a new opportunity for hackers to invade public privacy. With these risks in minds, users question the necessity of installing software on their smartphones that they don’t need to rely on or use every day. The potential threats are summarized below:

- Low awareness and limited motivation
- Limited function in two-way communication
- Unequal access limited data types
- Limited capacity in data management and sharing
- Limited data mining for citizen-reported data
- Lack lasting investment and strategic management
- Information security and privacy issues

**Discussion**

Based on the findings from the selected cases, this paper suggests a technical development framework that can be used to improve the existing mobile-based environmental information system.

**Technical Development Framework**

Understanding users’ needs is essential for using VGI methods in the environmental management field. In general, there are two kinds of needs. One is an urgent need, peak values, such as a disaster or a hazard. The other is a daily need, such as reading the news every day or checking the weather every day. When designing mobile information platforms, designers also need to think about what kinds of needs the public has. System deployment includes two tasks: front-end tasks and back-end tasks. Front ends represent the client side of the system, such as desktop users and mobile users. Back ends are the
server side of the system. It includes two web servers and a GIS server. There are two types of front ends: web front ends and mobile front ends. Web front ends are web applications that allow Internet clients or users to request back-end services through a URL (Uniform Resource Locator) via web browsers. Mobile front-ends are mobile applications that allow mobile clients to request back-end services through smartphones or tablets. Mobile applications can be downloaded and installed via Apple Store or Google Play. Typically, multi-requests occur when there are too many front-end users accessing the server simultaneously. The number of multi-requests from mobile clients is usually higher than the number of multi-requests from desktop clients, which means that the server back-ends should be optimized to account for these differences. Besides the REST (REST, Representational state transfer) service hosted on a GIS server, there is also educational information hosted on the back end. Educational information is typically placed on static web pages which does not require communication with a GIS server to access. By adding an additional web server (Figure 4), the system can filter the client into two groups: those requiring use of GIS services and those who do not. Using this deployment method, the total number of multi-requests for a GIS server can be reduced significantly if a certain number of clients only browse the education web page and, at the same time, those who search on the map or report data can still have a smooth experience. In short, this deployment method can reduce the peak number of multi-requests from clients, especially when there are many mobile users who are only viewing static information through the front end. A workstation was also added to the back end in order to give data access to professional users such as GIS analysts (Figure 1). In addition, by publishing web maps through ArcGIS online.com for related research, data can also be shared between other professionals such as environmental scientists and planners (Figure 1).

Figure 1: Deployment of front ends and back ends.
The key point of system architecture is using embedded Javascripts in Html (HyperText Markup Language) (Figure 2). Javascripts is a client-side scripting language. Using embedded Javascripts is an affordable way for deploying small business applications through clients, because most current browsers, such as Firefox, Safari or Chrome, support Javascripts very well. Thus, it is possible to develop only one application and make it executable on different platforms. The PhoneGap (PhoneGap, a mobile development framework produced by Nitobi) has realized this possibility; developers do not need to update their code frequently or develop applications for each different platform, which can significantly reduce the total cost of the system. The Html code with embedded Javascripts can be wrapped into the native environment on the client side, because PhoneGap provides a bridge between the native environment and web environment (Figure 2).

### System Architecture

![System Architecture Diagram](image)

**iOS/Android Native Environment**

<table>
<thead>
<tr>
<th>PhoneGap</th>
</tr>
</thead>
<tbody>
<tr>
<td>jQuery Mobile Widgets/html5</td>
</tr>
<tr>
<td>Embedded Javascripts</td>
</tr>
<tr>
<td>REST API (Application Programming Interface)</td>
</tr>
</tbody>
</table>

**Figure 2: Architecture of front ends and back ends.**

### Mobile front-end design framework

The design framework for mobile front ends, including visualization and user interfaces, includes five key features (Figure 3): (1) GPS reporting and mapping features will enable users to browse maps, send geocoded photos or data, and query attributes of geometric objects. (2) Publication and education features will be designed for novice technicians or students to study green infrastructure strategies. Third party publications are also posted as a link in this feature. (3) With the linkage to social media, users, experts and advocates can share their ideas through social networks. Several popular social networks will be included, such as Twitter and Facebook. (4) News and research progress will be posted through a research exhibition feature for environmental experts. (5) Users can also find reviews and contact information through feedback and contacts features.
Native app vs hybrid app vs web app

A mobile application, the front end of the VGI system, has a great impact on attracting volunteers and promoting VGI concepts to the public. Issues to consider include whether it appears user-friendly, and whether the key functionality enhances the user experience. There are three different types of mobile applications: native apps, hybrid apps, and web apps. All of these applications have their advantages and disadvantages when building a VGI system. Developers and planners also need to choose a suitable application type for their projects. It is difficult to assess which type is the perfect option for building a VGI system; planners and developers need to balance development costs and time, as well as key functional features.

Native applications are those that are developed by native programming languages, which are specified by mobile operating systems (OS). A native application can only run on the mobile OS that supports it. One immediate advantage of a native application is that all the documentation can be done with or without Internet connectivity (Robert, 2014). A native application can also run smoothly on the specified OS, and it usually has fewer bugs. However, a native application also has disadvantages. Since the native programming language is used to develop native apps, it is hard for a native app to be cross-platform. If planners or developers choose to develop a native application, they need to program on every OS platform by using different native OS coding languages, such as Objective-C and Java (Figure 4). In addition, updating the native app is also a problem since it requires knowledge of different programming languages. It can be expensive and time-consuming to develop a native app for the VGI concept, but in some cases, such as local data analysis, choosing a native application is a wise solution.
Web applications actually are websites in mobile form, which rely on browser and Internet connectivity. Web applications don’t have a native look, but they are much cheaper to develop than native apps and hybrid apps. Web apps are not popular, because they require users to remember website links. In general, web apps are not suitable for promoting VGI concepts because they don’t look user-friendly, and they often has a blurred user experience. Hybrid apps are part native apps, and part web apps. Hybrid app development requires web experience, such as Html5 and JavaScript. Although hybrid apps are not developed using native programming languages, they can have a native look and can access the functionalities of the smart phone. The native appearance of a hybrid app relies on some open sourced or third party user interface framework, such as jQuery Mobile and Sencha Touch. Thus, its performance may not be as smooth as a native app, but the cost and maintenance of the hybrid app is usually cheaper and easier than the native app workflow. In addition, choosing a hybrid app doesn’t require planners and developers to have native development experience; it only requires web development experience, making it easier for a hybrid app to be cross-platform. Most social media apps are hybrid apps, because they are easily distributed on different mobile OS platforms. In general, choosing the right type of app for building a VGI system is very important, because it has a direct impact on the volunteers. It is hard to say which type of app workflow is better than others. All have their own advantages and disadvantages.
Conclusions

Environmental programs with a relatively simple task and a large temporal and spatial extent will have more opportunities to be successful. A successful big data translation platform should provide a simple and easy-to-use platform to attract more public users. This platform can increase public awareness of environmental issues as well as collect community-submitted data. However, this new technology, which is a tool for facilitating public participation in environmental management processes, also has weaknesses, i.e., data quality issues. At the same time, the participation of citizens in collecting big environmental data is threatened by their lack of motivation, failed two-way communication, unequal access, limited data types, and ignorance of sustainable investment and management opportunities. Various opportunities, including integration with traditional communication techniques and social media, will enable more efficient and ubiquitous use of big environmental datasets in urban environmental management. Although weaknesses and threats such as data quality issues do exist in these cases, the strengths and opportunities from these case studies support the encouragement of the integration of citizen-participated environmental information into urban environmental management.

This paper provides suggestions for better use of this tool in the environmental management field. In general, the VGI method is not expected to replace traditional methods, but rather be an alternative solution for environmental monitoring and management. The data collected by non-professionals from a VGI system is reliable, and can be an additional data source for environmental monitoring and management.

Although the concept of VGI is not new, it has been applied to solving societal issues for at least 200 years. Cutting edge web technology makes it possible to implement VGI on a widespread basis. Data obtained from VGI methods, similar to crowdsourcing, provide invaluable results for decision-making. Like many advanced technologies created by human society, VGI concepts and methods were adopted by the community, regional, and urban planning fields only after successful use in the environmental domain. VGI methods have been used for monitoring the environment since 2007, and have realized many benefits such as cost savings for environmental projects and additional datasets for research. From the perspective of government agencies or authorized information organizations, VGI can compensate for the weaknesses of traditional mapping and remote sensing methods, saving the costs of mapping and maintaining the databases, breaking through the spatial limitations of information availability, and helping to promote environmental awareness. Bringing public engagement into environmental management has been a critical issue due to the gaps that exist between the environment, citizens, and governments. Other VGI practices are included in this research for exploring VGI methods in the planning field. The problems and concerns that we have confronted during this exploration include data quality, technology, user interface design issues, and other related factors. It has been proven that the data generated by VGI methods is believable if a large group of people participate in the data collection; however, the lack of motivation can decrease data quality if the VGI content or ideas cannot meet the public's needs. When utilizing VGI methods, planners and developers also need to understand public needs.

A variety of workflows were also tested in our research. There are three development workflows for building VGI mobile application. Native apps and hybrid apps
are suitable for developing the client side of a VGI system; both have their own advantages and disadvantages. Choosing which technology workflow to utilize depends on the project requirements as well as the professional skills of planners and developers. Finally, when utilizing VGI methods, planners need to think about other concerns, such as the value of the contributions from the public, how to encourage the government, the marketing aspects of mobile apps, and all of the factors and knowledge that could be helpful to promote public involvement.

Acknowledgements
This paper has been funded wholly or partially by the United States Environmental Protection Agency (EPA) under an assistance agreement (CD-977422010; UW-97735101). The contents do not necessarily reflect the views and policies of the EPA, nor does the mention of trade names or commercial products constitute endorsement or recommendation for use.

References


Fast food data: Where user-generated content works and where it doesn’t

Robert Manduca¹, Seth E. Spielman², and David Folch²

¹Department of Urban Studies and Planning, MIT
²Department of Geography, University of Colorado

July 15, 2014
Abstract

From a data collection perspective, user-generated “big” urban data offers clear advantages over traditional data sources: 1) it is available in real time with very short update cycles and 2) it is inexpensive to collect, in large part because it is often a byproduct of modern urban life. However, from a data use perspective the arguments for big urban data are much less compelling. There are real questions about the fidelity of big urban data to on the ground conditions. What phenomena are well captured by big urban data and what phenomena are not?

In this paper we examine the correspondence between Yelp data and an administrative dataset of restaurants in the Phoenix, Arizona metro area produced by the Maricopa Association of Governments (MAG). We find that the two datasets capture largely disjoint subsets of Phoenix restaurants, with only about one third of restaurants in each dataset present in the other. Point pattern analysis indicates that the Yelp data is significantly clustered relative to the MAG data. Specifically, restaurants in the Yelp data are concentrated in certain parts of metro Phoenix, most notably the downtowns of Phoenix, Scottsdale, and Tempe. Further analysis using US Census Workplace Area Characteristics data indicates that areas with more Yelp than MAG restaurants tend to have large numbers of college-educated workers and workers employed in the Arts, Entertainment, and Recreation sector.

Our comparison of Yelp and MAG data highlights the strengths and weaknesses of each. The Yelp data is far more detailed and comprehensive in certain areas of Phoenix, while the MAG data is likely to be more consistent across the entire region due to its systematic construction. When combined, administrative and user generated databases seem to provide a more holistic and comprehensive picture of the world than either would provide by itself.
1 Introduction

The promise of “big” urban data is that it provides a new way to understand and manage cities. Sensors embedded on buses, buildings, sewers, and roads can highlight the flow of energy and material within a city. Residents of a city, whether permanent or just passing through, can actively (or passively) serve as sensors by contributing information about the urban landscape and urban processes. Big urban data, in all its forms, offers several clear advantages over traditional administrative or survey data sources: 1) It is widely available in real time and has very short update cycles, 2) it is inexpensive to collect in large part because it often a byproduct of modern urban life, and 3) it describes urban processes that are not readily visible/available from traditional data collection mechanisms. In a sense big urban data is like fast food: cheap to acquire, ubiquitous, and always fresh from the oven (or at least freshly defrosted and deep fried). While the advantages of big urban data are clear and compelling in theory, in practice the utility of big urban data is much less clear. The uses of big data are limited in part because it is unclear exactly what big data is (and is not) telling us about a city. In this paper we engage this question through an analysis of the locations of restaurants in the Phoenix, AZ metropolitan area.

Big data is a very vague term. In this paper we focus on a narrow subset of what one might call big data, user generated content from the Yelp academic dataset\(^1\). We compare the Yelp data to a more traditional administrative source of data from the Maricopa Council of Governments (MAG). Maricopa County is part of the Phoenix, AZ metropolitan area. In line with our analogy between big data and fast food, we examine differences in the spatial distribution of restaurants between the two data sets. At the outset of this analysis we assumed that the MAG official register of all licensed restaurants would represent reality and that Yelp would capture only some aspect of the “truth.” Instead we found that both administrative and “big” urban data sets have omissions. The Yelp and the MAG data have roughly the same number of restaurants (4803 and 4181, respectively). However, the overlap between the two is relatively small: we could only match about a third of restaurants across datasets. These omissions are systematic and predictable. In certain kinds of places the Yelp data set seem to provide better coverage of restaurants than the official MAG database. Areas with a high density of Yelp relative to MAG restaurants tend to have more college-educated workers and more workers in the arts, entertainment, and recreation. These characteristics, combined with qualitative knowledge of the neighborhoods in question, begin to suggest that the Yelp data does an especially good job of documenting dense, hip areas, frequented by professionals and the workers who serve them. MAGs systematically collected data, on the other hand, is less exhaustive in these areas but is perhaps more complete in other parts of the metro area. Additionally, as described

\(^1\)Available from: [https://www.yelp.com/academic_dataset](https://www.yelp.com/academic_dataset)
in later sections, we use the US Census’ Longitudinal Employer Household Dynamics Database\(^2\) to use high-resolution (block-level) economic data to model the difference between the Yelp and MAG datasets with a fairly high level of accuracy ($r^2 = 0.72$).

This analysis highlights the primary challenges to the use of big urban data: While the data is current, the extent of its coverage is often unknown, and gaps in the data could be due to an absence of whatever is being measured or to an absence of “measurers,” people/instruments that upload data. That is, in user generated content citizens serve as both sensors and censors. There are real questions about the fidelity of big urban data to “real” on the ground conditions. What phenomena are well captured by big urban data and what phenomena are not? Here we offer an answer to this question for a particular type of phenomena (restaurants), with a particular type of data (Yelp) in a particular place (Phoenix, AZ).

2 Background

Because big urban data sets are not designed studies in formal statistical sense it is difficult to measure data quality. Thus much of the literature on “big” geographic has focuses on epistemological questions about what constitutes “good” user-generated content (Spielman, 2014). There are those who aim to formally evaluate the spatial accuracy of volunteered information (for example Haklay (Haklay, 2010)). These formal assessments are rooted in the concept of “spatial error” (Chrisman, 1991) that considers the positional accuracy of geographic data, the accuracy of attributes, the completeness of the data, and other factors that with sufficient resources and time can be objectively assessed. Goodchild and Li suggest that the quality of individual contributions can be checked for accuracy against a set of rules that define plausible arrangements of geographic features—for example the physical properties of a user-contributed coastline can be checked to ensure that they have a fractal structure as identified by Mandelbrot (Goodchild and Li, 2012; Mandelbrot, 1967). Mooney et al. take a similar approach to validation, arguing that new contributions should be geometrically similar to existing contributions of the same type, e.g. new lakes should have the same approximate geometric properties as existing lakes (Mooney et al., 2010). Others argue that discrepancies in volunteered information may not constitute errors (Elwood, 2008; Elwood et al., 2012). Different individuals can have different perspectives on the geographic world and the validity of these perspectives, and hence the validity of volunteered geographic information, can be judged based upon the credibility of the volunteer (Flanagan and Metzger, 2008). In this perspective it is difficult to objectively assess the accuracy of the data but one can assess its credibility via the reputation,\(^3\)

\(^2\)Available from: [http://lehd.ces.census.gov/data/](http://lehd.ces.census.gov/data/)
trustworthiness, and motivations of the contributor (?).

Evaluations of data are further complicated when data quality is not the primary goal of the data collection process (Arribas-Bel, 2014). When people share photos, opinions, and experiences on social media, spatial data is collected as a byproduct. In these circumstances, the nature of the primary action becomes important for proper interpretation of the gathered data (Cramer et al., 2011; Lindqvist et al., 2011). For example, restaurants are often well represented on check-in services such as Foursquare, while arts and entertainment destinations figure more prominently on story sharing services such as storyplace.me (Bentley et al., 2014). Additionally, users sometimes will use the location field for communicative purposes, such as humorous banter with friends for example checking in at a friends house but claiming that it is a strip club (Rost et al., 2013). In the case of review websites, contributions may be more likely among those who had particularly good or bad experiences (Brandes et al., 2013). Additionally, the representativeness of the contributing population is an issue in both volunteered and incidental geographic information (Crawford, 2011). Most user-generated datasets are not created by a subset of the population, and failing to account for that may lead to biased conclusions (Tufekci, 2013).

Because of these concerns, many recent studies utilizing incidental geographic data have attempted to validate their results through comparison with the results of more traditional analysis. Some studies have done this qualitatively, through interviews with stakeholders (Cranshaw et al., 2012). A more common approach is comparison with quantitative data, often gathered administratively at lower resolution than the incidental data or for only a portion of the full dataset. Examples of this approach applied to social media include the validation of geotagged Flickr photos as a measure of visitation to tourist attractions by comparing it to empirical visitor surveys from each site (Wood et al., 2013), the validation of using Twitter to measure mobility patterns by comparing it with tourist arrivals by country (Hawelka et al., 2014), the comparison of population estimates generated using Foursquare data to rasterized census data (Aubrecht et al., 2011).

Incidental geographic information isn’t limited to social media. A prime example of ancillary location data being repurposed for social science is the use of cell phone traces to study mobility patterns. Cell phone data has been embraced by the transportation community for its high temporal resolution and extensive coverage compared to traditional travel surveys (Caceres, 2007; Ratti et al., 2006; Zhang et al., 2010). As part of this adoption, cell phone data has been compared to traditional survey data in cases where both are available and found to exhibit similar patterns (Schneider et al., 2013; Calabrese et al., 2011; Jiang et al., 2013). Similarly, attempts to utilize requests to city Community Relationship Management systems for services such as graffiti removal or sidewalk repair as a measure of physical disorder have been validated through comparisons with
visual audits of the neighborhoods in question (O’Brien et al., 2013).

Finally, administrative datasets have been used for training rather than validation. Rodrigues et al. use data from administrative data sources such as Dun and Bradstreet to train classification algorithms on Yahoo! point of interest data (Rodrigues et al., 2013). Notably, they found relatively little overlap between the POIs in the Yahoo! data and those in the administrative datasets.

3 Methods and Materials

3.1 Data Sources and Processing

We use two primary data sources, The Yelp Academic Data Set, downloaded on March 26, 2013, and the MAG Employer Database. The Yelp Academic Data Set contains all Yelp reviews for Maricopa County, Arizona. Based on data from Dun and Bradstreet and augmented manually by member agencies, the MAG dataset contains information on all establishments with at least five registered employees. Both datasets contained the establishment name, address, and latitude and longitude. The Yelp dataset includes information on reviews and reviewers, which is not utilized in this analysis. Establishments in the Yelp database are tagged with one or more codes describing the type of business, which loosely correspond to the standard industry classification codes (NAICS) widely used in other databases. We selected all establishments coded as “Restaurants”, “Food Retailers”, or “Nightlife,” then dropped those flagged as supermarkets, liquor stores, convenience stores, caterers, food trucks, and farmers markets. We then developed a crosswalk between the NAICS codes and Yelp classification codes and selected businesses from the MAG file using the 3 digit NAICS codes that corresponded to the Yelp categories. This selection procedure was imperfect; for example, the three digit NAICS code “722” includes both restaurants and “gentlemen’s clubs” in the MAG database. We further refined the selection by deleting specific six NAICS codes from the MAG file.

Furthermore, there were some inconsistencies in geographic extent between the files. Yelp included businesses outside the jurisdiction of the Maricopa Council of Government; these were deleted. Both data sets included a few small exurban towns separated from the main parts of Phoenix by miles of undeveloped land, which were deleted as well. The final restaurants included

3From the initial selection of all businesses with NAICS codes 722 and 445 we deleted codes: 445110 (supermarkets), 445120 (convenience stores), 45310 (liquor stores), and 722320 (caterers) to keep our analysis focused on restaurants, the traditional focus of Yelp reviews.

4From Yelp we deleted the towns of San Tan Valley, Apache Junction, Casa Grande, Maricopa, Florence, Coolidge, Tortilla Flat, Gold Canyon, Tonopah, Gila Bend, Wickenburg, and Morristown. From MAG we deleted Wittmann, New River, Tonopah, Tortilla Flat, Apache Junction, Wickenburg, Morristown, and Gila Bend
in our analysis are mapped in Figure 1.

3.2 Matching

We tried to link restaurants from Yelp to MAG file (and vice versa) using a variety of approaches. To begin, we geocoded all establishments using the Google Geocoding API. This returned a clean address field in a consistent format. We first attempted to link the files by blocking location, that is for any establishment the set of potential matches is set of restaurants geocoding to the same location. However, differences in addressing between the two files made this blocking strategy too restrictive, thus we relaxed the co-location criteria selecting as potential matches any restaurant pairs found less than 3000 feet apart. Co-located files are not necessarily matches as many strip malls and food courts geocode to the same location. With our set of potential matches, we conducted a “fuzzy match” on the establishment name using the FuzzyWuzzy package in Python (SeatGeek, 2014). This approach is used to match strings that are mostly but not entirely identical, and is a common way to account for misspellings and other difficulties inherent in working with
language. Potential matches that had a sufficiently high fuzzy match score were deemed to be the same restaurant. Thus, we considered restaurant pairs to be matches if they either had the same address in both datasets or were found less than 3000 feet apart, and they had names that were sufficiently similar.

After conducting the automated matching process, we manually checked the potential address matches. This turned up roughly 200 additional matches, mostly instances where an abbreviation was used in one dataset but not the other (e.g. “Kentucky Fried Chicken” vs. “KFC”), or where the noun used differed in the two datasets (“Christys Grill” vs “Christys Restaurant”). To our surprise, this matching procedure yielded a set of only 1726 matches out of over 5000 restaurants in each dataset. Further investigation suggested that while there were some cases of errors or out-of-date records in each file, the majority of the unmatched restaurants in both files were in fact real and currently open. In an effort to understand the discrepancy between the two files we conducted a more detailed spatial/statistical analysis.

### 3.3 Spatial Analysis

Given the difficulties linking the two files, further data processing was necessary to conduct the spatial analysis. We converted the address-level (point) data files into gridded maps of restaurant density using the Spatstat package in R \cite{Baddeley2005}. Kernel based maps of restaurant density (restaurants per square km) and maps of the absolute number (count) of restaurants per pixel were produced. Grid based (raster) maps were created at two resolutions, a fine resolution which divided the study area into a $128 \times 128$ pixel array yielding pixels that were 791m by 641m and a coarse resolution which divided the study area into a $32 \times 32$ pixel array yielding pixels of 3160m by 2560m. These resolutions were selected because we felt that using higher resolution would prevent nearby or co-located but mis-geocoded restaurants from being in the same pixel. Using census units, such as blocks or tracts, as a unit of analysis was problematic because many restaurants in Phoenix are located at intersections of major roads, which also happen to be boundaries between tracts, thus restaurants at the same intersection we being placed into different census units. To calculate the Kernel density we estimated the optimal bandwidth using a likelihood cross validation method implemented in Spatstat and attributed to Loader \cite{Loader1999}. In addition to creating density surfaces for all restaurants, Yelp restaurants, and MAG restaurants, we also computed a surface of the difference in density between Yelp and MAG. These maps are shown in Figure 2.
Figure 2. Restaurant Densities
3.4 Point Pattern Analysis

On the basis of visual inspection (see figure 1), it is difficult to determine the differences in the spatial distribution of the MAG and Yelp points. To explore these distributions, we undertook a formal point pattern analysis. Point pattern analysis is a set of methods for developing statistical hypothesis tests about the spatial distribution of “events.” The use of the word “event” to describe a discrete location in space hints at an essential element of point pattern analysis. In point pattern analysis, one examines a set of observed locations relative to some probability surface describing the potential location of events. For example, if a lightning strike was equally probable at any location within a study area, the probability surface describing the probability of lightning strikes would be uniform. However, if hilltops had a higher probability of being struck by lightning then the probability would be non-uniform, or in the language of point pattern analysis “inhomogeneous.” The probability surface used in point pattern analysis is critical to the inference drawn from the analysis. Most analyses use a uniform Poisson probability distribution, which suggests that the number of events within an area follows a Poisson distribution that is the same in the entire study area (this is sometimes called Complete Spatial Randomness (CSR)). One does not need to conduct a statistical test to know that restaurants are not randomly located within metropolitan Phoenix.

We conducted a series of progressively sophisticated point pattern analyses; here we only describe the most interesting analysis. First, we developed a probability surface that said there is zero probability of restaurants locating in an area that has fewer than 0.1 restaurants per square kilometer. This removed parks and residential areas from the point pattern analysis. We then computed separate L-functions, a variance stabilizing version of Ripley’s K-Function, for both the MAG and the Yelp data, \( L_{mag}(f) \) and \( L_{yelp}(f) \) respectively. The L-functions are essentially a curve, estimated from the observed data, which describes the number of restaurants near an arbitrarily selected restaurant as a function of distance. If two point patterns have the same spatial pattern, the difference between their L-functions should be approximately zero. That is, if the number of Yelp restaurants within 100m, 200m, 300m, 500m, . . . , 10000m of an arbitrarily selected Yelp restaurant is approximately the same as the number of MAG restaurants near a MAG restaurant at each distance band, \( L_{yelp}(f) - L_{mag}(f) = 0 \), this would imply that the points had the same spatial pattern. We tested the null hypothesis \( H_0 : L_{yelp}(f) = L_{mag}(f) \) using the Diggle-Chetwynd Test (Diggle and Chetwynd [1991]) and rejected the null hypothesis (using 99 simulations, p-value .01). This indicates that the Yelp and MAG data have different spatial distributions; in general, Yelp restaurants are more closely packed at some distance bands than MAG restaurants.

Next we tested the null hypothesis that Yelp restaurants are located randomly with respect to
MAG restaurants using a cross k-function. This null hypothesis differs from the first in that it assesses the independence of the two point patterns. Here we found, using a hypothesis testing strategy from Loosmore and Ford (?), that location of Yelp restaurants is not independent from the location of the MAG restaurants (p-value 0.01 using 99 simulations of random Yelp locations with respect to MAG). At some scales the number of Yelp restaurants near a MAG restaurant exceeds the simulation envelope, suggesting an excess of Yelp restaurants near MAG restaurants. In this preliminary analysis we do not have the space to fully enumerate the point pattern hypothesis tests, more detail is available from the authors.

3.5 Models

The Point pattern analysis suggests that Yelp restaurants are clustered around MAG restaurants. To further explore this idea we developed a series of statistical models, the first model $Y$ is the number of Yelp restaurants in a pixel and $x$ is the number of MAG restaurants in a pixel. The model $Y = -0.8 + 1.28x$ has an $r^2 = 0.78$ and the model $Y = 0.3 + 0.9x + 0.0002x^3$ has an adjusted $r^2 = 0.84$, suggesting that the number of Yelp restaurants in a pixel can be reasonably seen as an exponential function of the number of MAG restaurants. Stated in plain English, as the number of MAG restaurants increases the number of Yelp restaurants increases, but the rate of increase for Yelp is faster than the rate of increase for MAG. Areas with a lot of MAG restaurants have more Yelp restaurants.

In an effort to understand the characteristics of the places with developed a second set of models where $Y$ was the difference between the number of MAG and Yelp restaurants in each pixel. This model used a large database called the Work Place Area Characteristics File (WAC) from US census bureau. This WAC file contains a total of 51 variables that count the number jobs in a census block by industry, age of worker, age of firm, race of worker, education level of worker, and monthly wages. In line with the exploratory nature of this paper we uses an exhaustive regression search procedure that fit all possible 1-7 variable regression models. Just using the number of MAG restaurants gives an $r^2$ of 0.15. Including the number of jobs for workers with Bachelor’s degree or advanced degree, the most powerful single predictive variable, increases the $r^2$ to 0.43. Including the number of jobs in Arts, Entertainment, and Recreation further increases it to 0.47.

4 Results

Contrary to our initial expectations, we found that although the number of restaurants in the Yelp and MAG datasets is similar, the restaurants themselves are different: Only about one third of the
restaurants present in each dataset are also found in the other. Further, the spatial distribution of restaurants differs by dataset: as seen in Figure 2d, the concentration of Yelp restaurants is higher in central and northern Maricopa County, while that of MAG restaurants is higher in the southeast and northwest. Point pattern analysis indicated that this difference in spatial distribution is statistically significant, with Yelp restaurants being more clustered than MAG restaurants, but that Yelp restaurants are not located independently of MAG ones. Further investigation of the characteristics of areas with large numbers of Yelp restaurants relative to MAG ones found that these areas—the most prominent are downtown Phoenix, Scottsdale, and Tempe—tend to have larger numbers of college-educated workers and larger numbers of jobs in arts, entertainment, and recreation.

5 Discussion

Our comparison of Yelp and MAG data highlights the strengths and weaknesses of each. The Yelp data is able to be far more detailed and comprehensive in certain areas of Phoenix, providing information on a large number of restaurants. It also may be better at documenting informal or family-run businesses, and it certainly is updated more frequently than MAG. However, Yelp is not equally present across the metro area, so analysis of Yelp data alone would incorrectly suggest that restaurants are concentrated primarily in downtown areas. MAG data may be less exhaustive in certain places, but since it draws on administrative records it is likely to be more consistent across the entire city. An effective planner might use the more evenly spread MAG data for metro-level research, supplementing it with Yelp data for detailed analysis of the neighborhoods well served by Yelp.

The different strengths of the two datasets are just the most recent iteration of an established problem. Datasets on businesses have almost always been generated as the byproduct of some other, commercial process. The nature of that process influences the form of the final dataset. By combining features of multiple datasets generated through different processes it may be possible to address the weaknesses of each.

References


Cranshaw, J., R. Schwartz, J. Hong, and N. Sadeh (2012). The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. ICWSM.


Using Social Media Data to Understand Cities
Dan Tasse, Jason I. Hong
Human-Computer Interaction Institute
Carnegie Mellon University
dantasse@cmu.edu, jasonh@cs.cmu.edu
Abstract
Understanding urban dynamics is crucial for a number of domains, but it can be expensive and time consuming to gather necessary data. The rapid rise of social media has given us a new and massive source of geotagged data that can be transformative in terms of how we understand our cities. In this position paper, we describe three opportunities in using geotagged social media data: to help city planners, to help small businesses, and to help individuals adapt to their city better. We also sketch some possible research projects to help map out the design space, as well as discuss some limitations and challenges in using this kind of data.

Introduction
Over half of the world’s population now lives in cities [33], and understanding the cities we live in has never been more important. Urban planners need to plan future developments, transit authorities need to optimize routes, and people need to effectively integrate into their communities.

Currently, a number of methods are used to collect data about people, but these methods tend to be slow, labor-intensive, expensive, and lead to relatively sparse data. For example, the US census cost $13 billion in 2010 [11], and is only collected once every ten years. The American Community Survey is collected annually, and cost about $170 million in 2012, but only samples around 1% of households in any given year [19]. While data like this can benefit planners, policy makers, researchers, and businesses in understanding changes over time and how to allocate resources, today’s methods for understanding people and cities are slow, expensive, labor-intensive, and do not scale well.

Researchers have looked at using proprietary call detail records (CDRs) from telecoms to model mobility patterns [3, 18, 24] and other social patterns, such as the size of one’s social network and one’s relationship with others [36]. These studies leverage millions of data points; however, these approaches also have coarse location granularity (up to 1 sq. mile), are somewhat sparse (CDRs are recorded only when a call or SMS is made), have minimal context (location, date, caller, callee), and use data not generally available to others. Similarly, researchers have also looked at having participants install custom apps. However, this approach has challenges in scaling up to cities, given the large number of app users needed to get useful data. Corporations have also surreptitiously installed software on people’s smartphones (such as CallerIQ [34] and Verizon’s Precision Market Insights [35]), though this has led to widespread outcry due to privacy concerns.

We argue that there is an exciting opportunity for creating new ways to conceptualize and visualize the dynamics, structure, and character of a city by analyzing the social media its residents already generate. Millions of people already use Twitter, Instagram, Foursquare, and other social media services to update their friends about where they are, communicate with friends and strangers, and record their actions. The sheer quantity of data is also tantalizing: Twitter claims that its
users send over 500 million tweets daily, and Instagram claims its users share about 60 million photos per day [23]. Some of this media is geotagged with GPS data, making it possible to start inferring people’s behaviors over time. In contrast to CDRs from telcos, we can get fine-grained location data, and at times beyond when people make phone calls. In contrast to having people install custom apps (which is hard to persuade people to do), we can leverage social media data that millions of people are already creating every day.

We believe that this kind of geotagged social media data, combined with new kinds of analytics tools, will let urban planners, policy analysts, social scientists, and computer scientists explore how people actually use a city, in a manner that is cheap, highly scalable, and insightful. These tools can shed light onto the factors that come together to shape the urban landscape and the social texture of city life, including municipal borders, demographics, economic development, resources, geography, and planning.

As such, our main question here is, how can we use this kind of publicly visible, geotagged social media data to help us understand cities better? In this position paper, we sketch out several opportunities for new kinds of analytics tools based on geotagged social media data. We also discuss some longer-term challenges in using this kind of data, including biases in this kind of data, issues of privacy, and fostering a sustainable ecosystem where the value of this kind of data is shared with more people.

Opportunities

In this section, we sketch out some design and research opportunities, looking at three specific application areas. Many of the ideas we discuss below are speculative. We use these ideas as a way of describing the potential of geotagged social media data, as well as offering possible new directions for the research community.

For City Planners

First, and perhaps most promisingly, we believe that geotagged social media data can offer city planners and developers better information that can be used to improve planning and quality of life in cities. This might include new kinds of metrics for understanding people’s interactions in different parts of a city, new methods of pinpointing problems that people are facing, and new ways of identifying potential opportunities for improving things.

Mapping Socioeconomic Status

It is important for governments to know the socioeconomic status of different sections of their jurisdiction in order to properly allocate resources. In England, for example, the government uses the Index of Multiple Deprivation (IMD) to measure where the problems of poverty are the most severe, and to therefore mitigate those effects. The IMD is based on surveys and other statistics collected by different areas.
of government. However, even in developed countries, surveys and statistics can be difficult and expensive to collect.

Past work with cell phone call logs suggests that it is possible to find valuable demographic information using communication records. For example, Eagle et al found that network diversity in phone calls correlated with the IMD [16]. A recent project by Smith-Clarke et al [41] explored the use of call logs to map poverty. Especially in developing countries, maps of poverty are often very coarse and out of date. This is not a simple problem of improving an already-good metric; data at different granularities can tell very different stories. For example, Figure 1 shows two different maps of socioeconomic data in the UK, one very coarse-grained and one fine-grained.

However, call log data, while more complete than surveys, still presents the limitations mentioned earlier: it is proprietary, coarse-grained, and lacking context and transparency. Much social media data, on the other hand, is and publicly visible and accessible to researchers. Different forms of social media data also offer their own advantages. For example, Twitter users follow other users, and Foursquare checkins often have “likes” or comments attached.

![Figure 1](image)

(a) 32844 LSOAs, 2010  
(b) 9 regions, 2001

Figure 1 - Left: accurate data about socio-economic deprivation in England; darker indicates more deprivation. Right: much coarser, out-of-date information about the same index. Figure from [41].

**Mapping Quality of Life**

Socioeconomic status, however, is not the only metric that matters. A community can be poor but flourishing, or rich but suffering. Other metrics like violence, pollution, location efficiency, and even community coherence are important for cities to track. Some of these are even more difficult to track than socioeconomic status.
We believe some aspects of quality of life can be modeled using geotagged social media data. For example, approximating violence may be possible by analyzing the content of posts. Choudhury et al showed that psychological features associated with desensitization appeared over time in tweets by people affected by the Mexican Drug War [10]. Other work has found that sentiments in tweets are correlated with general socio-economic wellbeing [39]. Measuring and mapping posts that contain these emotional words may help us find high crime areas and measure the change over time.

As another example, location efficiency, or the total cost of transportation for someone living in a certain location, can be approximated by sites like Walkscore.com. However, Walkscore currently only relies on the spaces, that is, where services are on the map. It does not take into account the places, the ways that people use these services. A small market may be classified as a “grocery store”, but if nobody goes to it for groceries, maybe it actually fails to meet people’s grocery needs. We believe geotagged social media data can be used as a new way of understanding how people actually use places, and thereby offer a better measure of location efficiency.

**Mapping Mobility**

One more analysis that could be useful for city planners is in understanding the mobility patterns of people in different parts of different cities. This kind of information can help, for example, in planning transportation networks [27]. Mobility can help planners with social information as well, such as how public or private people feel a place is [42].

Previously, mobility information has been gathered from many sources, but they all lack the granularity and ease of collection of social media data. Cell tower data has been used to estimate the daily ranges of cell phone users [1]. At a larger scale, data from moving dollar bills has been used to understand the range of human travel [5]. Among other applications, this data could be used for economic purposes, such as understanding the value of centralized business districts like the Garment District in New York [43]. It seems plausible that pre-existing social media data could help us find the similar information without needing people to enter dollar bill serial numbers or phone companies to grant access to expensive and sensitive call logs. Geotagged social media data is also more fine-grained, allowing us to pinpoint specific venues that people are going to.

**“Design Patterns” for Cities**

Originating in the field of architecture, *design patterns* are good and reusable solutions to common design problems. Geotagged social media data offers new ways of analyzing physical spaces and understanding how the design of those spaces influences people’s behaviors.

For example, in his book *A Pattern Language* [1], Alexander and colleagues present several kinds of patterns characterizing communities and neighborhoods. These
patterns include Activity Nodes (community facilities should not be scattered individually through a city, but rather clustered together), Promenades (a center for its public life, a place to see people and to be seen), Shopping Streets (shopping centers should be located near major traffic arteries, but should be quiet and comfortable for pedestrians), and Night Life (places that are open late at night should be clustered together).

By analyzing geotagged social media data, we believe it is possible to extract known design patterns. One possible scenario is letting people search for design patterns in a given city, e.g. “where is Night Life in this city?” or “show the major Promenades”. Another possibility is to compare the relationship of different patterns in different cities, as a way of analyzing why certain designs work well and others do not. For example, one might find that areas that serve both as Shopping Streets and as Night Life are well correlated with vibrant communities and general well-being.

**For Small Businesses**

Understanding one’s customers is crucial for owners of small businesses, like restaurants, bars, and coffee shops. We envision two possible scenarios for how geotagged social media data can help small business owners.

**Knowing Demographics of Existing Customers**

Small businesses cannot easily compete with big-box stores in terms of data and analytics about existing customers. This makes it difficult for small businesses to tailor their services and advertisements effectively.

Businesses can already check their reviews on Yelp or Foursquare. We believe that geotagged social media data can offer different kinds of insights about the behaviors and demographics of customers. One example would be knowing what people do before and after visiting a given venue. For example, if a coffee shop owner finds that many people go to a sandwich shop after the coffee shop, they may want to partner with those kinds of stores or offer sandwiches themselves. This same analysis could be done with classes of venues, for example, cafés or donut shops.

As another example, an owner may want to do retail trade analysis [22], which is a kind of marketing research for understanding where a store's customers are coming from, how many potential customers are in a given area, and where one can look for more potential customers. Some examples include quantifying and visualizing the flow and movement of customers in the area around a given store. Using this kind of analysis, a business can select potential store locations, identify likely competitors, and pinpoint ideal places for advertisements.

Currently, retail trade analysis is labor intensive, consisting of numerous observations by field workers (e.g. watching where customers come from and where they go, or shadowing customers) or surveys given to customers. Publicly visible social media data offers a way of scaling up this kind of process, and
extending the kind of analysis beyond just the immediate area. For example, one could analyze more general kinds of patterns. For example, what are the most popular stores in this area, and how does the store in question stack up? How does the store in question compare against competitors in the same city?

Knowing more about the people themselves would be useful as well. For example, in general, what kinds of venues are most popular for people who come to this store? If a business finds that all of its patrons come from neighborhoods where live music is popular, they may want to consider hosting musicians themselves. All of the information offered by a service like this would have to be rather coarse, but it could provide new kinds of insights for small businesses.

Knowing Where to Locate a New Business
New businesses often have many different potential locations, and evaluating them can be difficult. Public social media data could give these business owners more insight into advantages and disadvantages of their potential sites. For example, if they find that a certain neighborhood has many people who visit Thai restaurants in other parts of the city, they could locate a new Thai restaurant there.

For Individuals
There are also many opportunities for using geotagged social media to benefit individuals as well. Below, we sketch out a few themes.

Feeling At Home In New Cities
Moving to a new city can make it hard for people to be part of a community. The formally defined boundaries of neighborhoods may help people understand the spaces where they live, but not so much the socially constructed places [20]. Currently it is difficult for non-locals to know the social constructs of a city as well as locals do. This is particularly important when someone changes places, either as a tourist or a new resident.

Imagine a new person arriving in a diverse city like San Francisco, with multiple neighborhoods and sub-neighborhoods. It would be useful for that person to know the types of people who live in the city, and where each group goes: students go to this neighborhood in the evening, members of the Italian community like to spend time in this area, families often live in this neighborhood but spend time in that neighborhood on weekends.

Some work has been done in this area, but we believe it could be extended. Komninos et al collected data from Foursquare to examine patterns over times of day and days of the week [28], showing the daily variations of people’s activity in a city in Greece. They showed when people are checking in, and at what kind of venue, but not who was checking in. Cheng et al, too, showed general patterns of mobility and times of checkins [9], but these statistics remain difficult for individuals to interpret.
Related, the Livehoods project [8] and Hoodsquare [45] both look at helping people understand their cities by clustering nearby places into neighborhoods. Livehoods used Foursquare checkins, clustering nearby places where the same people often checked in. Hoodsquare considered not only checkins but also other factors including time, location category, and whether tourists or locals attended the place. Both of these projects would be helpful for people to find their way in a new city, but even their output could be more informative. Instead of simply knowing that a neighborhood has certain boundaries, knowing why those boundaries are drawn or what people do inside those boundaries would be helpful. Andrienko et al [2] also describe a visual analytic approach to finding important places based on mobility data that could help newcomers understand which places were more popular or important.

Another approach to helping people get to know the city is to help them get to know the people in the city, rather than the places. We look to the work of Joseph et al [26] who used topic models to assign people to clusters such as “sports enthusiast” or “art enthusiast”. We could imagine this information being useful for individuals to find other like-minded people.

**Discovering New Places**
In the previous section, we described applications to help people become accustomed to a new city. However, sometimes the opposite problem may arise: people become too comfortable in their routines and they want to discover new places. Some recent projects have helped people to discover new places based on actions that can be performed there [13] or aspects that people love about the place [7]. However, we believe that this idea can be pushed further. Perhaps combining a person’s current mobility patterns with visualizations of other people’s mobility patterns would help a person to put their own actions in context. They may realize that there are entire social flows in the city that they did not even know existed.

**Understanding People, Not Just Places**
Tools like Yelp and Urbanspoon already exist to help people find places they would like to go or discover new places that they didn’t know about. Previous work like Livehoods [8] also worked to enable a richer understanding of the places there. The benefit of incorporating social media data, though, is that users can start to understand the people who go to places, not just the places themselves.

**Potential Research in this Design Space**
In this section we sketch out some potentially interesting research projects in using geotagged social media data. Our goal here is to map out different points in the overall design space, which can be useful in understanding the range of applications as well as the pros and cons of various techniques.

**Who Goes Here?**
The approach is simple: select a geographic region (which may be as small as an individual store) and retrieve all the tweets of the people who have ever tweeted
there. Then, compute a heat map or other geographic visualization of the other places that they tweet. This kind of visualization could help elucidate all the places associated with a given place, and could be useful to small business owners or managers of a larger organization like a university.

**Groceryshed**

Watershed maps show where water drains off to lakes and oceans. Researchers have extended this metaphor to map “laborsheds” and “paradesheds” [4] to describe where people who work in a certain area come from, or people who attend a certain parade. We could extend this metaphor even further to describe “sheds” of smaller categories of business, such as Thai restaurants.

More interestingly, we could map change over time in various “sheds”. This could be particularly important for grocery stores. Places that are outside any “groceryshed” could be candidate areas for a new store, and showing the change in people’s behavior after a new store goes in could help measure the impact of that store. The content of Tweets or other social data could also show how people’s behavior changed after a new grocery store was put in.

**How Is This Place Relevant To Me?**

We envision a system that can convey not only what a place is, but also what it means. Imagine a user looking up a particular coffee shop. Currently, they can look up the coffee shop’s web site, find basic information like store hours, and find reviews on sites like Yelp. Using geotagged social media data, however, we could surface information like:

- Your friends (or people you follow on Twitter) go here five times per week
- Friends of your friends go here much more than nearby coffee shops
- People who are music enthusiasts like you (using topic modeling as in [26]) often go to this coffee shop
- You’ve been to three other coffee shops that are very similar to this one
- People who tweet here show the same profiles of emotions as your tweets

These could help people form a deeper relationship with a place than one based on locality or business type alone. In addition, we could pre-compute measures of relevance for a particular user, giving them a map of places that they might enjoy.

**Human Network Visualizations**

We can go beyond assigning people to groups by topics, also showing where they tweet over time. This could help people understand the dynamics of neighborhoods where, for example, one group of more affluent people are pricing out a group of previous residents. One interesting work in this area is the Yelp word maps [44], which show where people write certain words, like “hipster”, in reviews of businesses. However, this still describes the places; using social media data, we could show maps that describe the people. Instead of a map of locations tagged as “hipster”, we could identify groups of people based on their check-in patterns and tag where they go during the day. Perhaps the hipsters frequent certain coffee shops
in the morning and certain bars at night, but during the day hang out in parks where they do not check in.

**Cheaper, Easier, and Richer Demographics**

For all of our groups and stakeholders, it is important to understand demographic information of city regions. We could improve the process in two main ways. First, we could make it cheaper and easier to infer existing demographic information. We plan to investigate whether Twitter volume, or volume of certain topics of discussion, correlates with deprivation or other measures of socioeconomic status or quality of life. If so, then we can use the social media measures as proxies for real measures, and thereby collect that information cheaply and in real time.

Second, we plan to create more descriptive demographics. In each neighborhood, we can calculate each person’s average distance traveled, radius of gyration, and other measures of mobility. We could either simply output statistics or we could create interactive visualizations that show the daily movements of people in each neighborhood.

**Data Sources and Limitations**

For our research, we are currently using data from Twitter, due to its richness and volume. Twitter claims over 500 million tweets are posted per day [29]. Furthermore, this data is publicly available. While only a small fraction of these tweets are geotagged, even a small fraction of tweets from any given day forms a large and rich data set. Furthermore, past work suggests that the sampling bias from only selecting these tweets is limited [38].

**Biases in the Data**

Of course, neither Twitter nor Foursquare provides an exactly accurate view of people’s mobility. Both are communicative media, not purely representative. For example, Rost et al [40] report that the Museum of Modern Art in New York has more check-ins than Atlanta’s airport, even though the airport had almost three times as many visitors in the time period that was studied. In some cases this will not matter; if we are clustering people, for example, grouping people who communicate that they go to the same places will be nearly as successful as grouping people who actually go to the same places. In other cases, we hope to minimize this bias by primarily comparing similar businesses, but we must remain aware of it.

Second, these media can be performative as well. People check in not because the check-in represents their location the most accurately, but because they want to show off that they have performed the check-in [6]. Sometimes people may avoid checking in for the same reason; they do not want it to be known that they checked in at a certain venue, like a fast food restaurant [30].
Third, there are currently several demographic biases in these data sets. For example, Twitter, Flickr, and Foursquare are all more active per capita in cities than outside them [21]. Furthermore, these social media sites are all used by predominantly young, male, technology-savvy people.

One effect of this bias is shown in Figure 2. This screenshot shows the results of some of our clusters in Livehoods [8], with each dot representing a venue, and different colors representing different clusters. Note the lack of data in the center of the figure. This area is Pittsburgh’s Hill District, a historic area which was the center of Pittsburgh’s jazz scene in the early 20th century. Currently, the median income for residents in the Hill District is far lower than in other parts of Pittsburgh. This neighborhood has also seen some revitalization with new senior housing, a library, a YMCA, several small office buildings, and a grocery store. However, there is still a notable lack of geotagged social media data in this area.

In short, while geotagged social media data has great potential, we also need to be careful because this data may not necessarily be representative of all people that live in a city. It is possible that this demographic bias may solve itself over time. Currently, about 58% of Americans have a smartphone [37], and the number is rapidly growing. However, it may still be many years before demographics are more representative, and there is still no guarantee that the demographics of geotagged
social media data will follow. For now, one approach is to look for ways of accounting for these kinds of biases in models. Another approach is to make clearer what the models do and do not represent.

**Privacy Implications**

Privacy is also a clear concern in using geotagged social media to understand cities. From an Institutional Review Board (IRB) perspective, much of social media data is considered exempt, because the researchers do not directly interact with participants, the data already exists, and the data is often publicly visible. However, as researchers, we need to go beyond IRB and offer stronger privacy protections, especially if we make our analyses available as interactive tools.

Here, there are at least two major privacy concerns. The first is making it easy to access detailed information about specific individuals. Even if a person’s social media data is public data, interactive tools could make a person’s history and inferences on that history more conspicuously available. Some trivial examples include algorithms for determining a user’s home and work locations based on their tweets [28]. More involved examples might include other aspects of their behaviors, such as their activities, preferences, and mobility patterns. In the Livehoods project, we mitigated this aspect of user privacy by only presenting information about locations, not people. We also removed all venues labeled as private homes.

Second, we need to be more careful and more thoughtful about the kinds of inferences that algorithms can make about people, as these inferences can have far-reaching effects, regardless of whether they are accurate or not. There are numerous examples of inferences outside of geotagged social media that might be viewed as intrusive, embarrassing, or even harmful. For example, Jernighan and Mistree [25] found that, given a social network with men who did not self-report their sexuality, they could identify gay men simply by analyzing the self-reported sexuality of an individual’s friends. As another example, the New York Times reported on how Target had developed algorithms that could infer if a customer was pregnant [14]. A separate New York Times article reported on how people were assessed for credit risks based on what they purchased as well as where they shopped [15].

It is important to note that these risks are not just hypothetical. At least one person had his credit card limit lowered, with the explanation that “other customers who have used their card at establishments where you recently shopped have a poor repayment history with American Express” [12].

It is not yet clear what the full range and extent of inferences is with geotagged social media. A significant concern is that inferences like the ones above and ones using social media data can become a proxy for socioeconomic status, gender, or race, inadvertently or even intentionally skirting around charged issues under the guise of an “objective” algorithm. It is also not clear if there is anything that can be done about these kinds of inferences, given that these inferences would be done on
private servers. It is unlikely that there is a technical solution to this problem. It may very well be the case that society will require new kinds of laws governing how these inferences are used, rather than trying to control the inferences themselves.

Creating a Sustainable Ecosystem
We hope to find a way to co-create value both to social media users and to people in our work. As it exists now, value flows only from users to marketers and analysts. To create a more sustainable tool, and to avoid impinging on users’ freedoms, it is important that the users gain some benefit from any system we create as well. Some of our projects point in this direction, especially the ones aimed at individual users. People may be more amenable to a tool that offers businesses insights based on their public tweets if they can have access to those insights as well.

A successful example of co-creation of value is Tiramisu [46]. This app aimed to provide real-time bus timing information by asking people to send messages to a server when they were on a bus. Users were allowed to get more information if they shared more information. In contrast, OneBusAway [17] provides real-time bus information using sensors that are installed on buses. Using a collaborative approach, it may not be necessary to implement a costly instrumentation project. In addition, people may feel more ownership of a system if they contribute to it.

Conclusion
Despite these challenges, social media remains a potentially transformative, yet untapped, source of geographic data. Analytics tools based on geotagged social media data can help city planners plan future developments, businesses understand the pulse of their customers, and individuals fit into new cities much more seamlessly. As our cities grow exponentially and more of the world moves into heavily urbanized areas, instead of using costly methods to understand our cities, researchers of all kinds will be able to mine existing data to understand social patterns that are already there.

References


Crowdsourcing Street Beauty: A New Method for Visual Preference Surveys

Draft prepared for Big Data and Urban Informatics Conference, August 11-12, 2014, Chicago, IL
Version Date: July 17, 2014

Robert Goodspeed
Taubman College of Architecture and Urban Planning
University of Michigan
2000 Bonisteel Boulevard
Ann Arbor, MI 48109
rgoodspe@umich.edu
1-734-615-7254

Jacob Yan
Taubman College of Architecture and Urban Planning
University of Michigan
jacobyan@umich.edu

Abstract

Measuring public visual preferences for streets and buildings is a perennial issue in professional planning. This paper reports a visual preference survey of street beauty completed online using crowdsourcing methods. Following Ewing et al. (2005), the resulting ratings were decomposed to environmental features using regression analysis. The paper demonstrates a method with several novel features: the use of crowdsourcing resulting in thousands of voters and over 100,000 separate votes, the use of an Elo rating algorithm to compute a beauty index, and the combination of preference analysis with regression and the use of GIS data to construct environmental indicators. The analysis finds ten variables can explain roughly 40% of the variation in the beauty index. The most important predictors are the number of trees, number of intersections in the vicinity of the observation point. In addition, historic buildings, the presence of parks and open spaces, and bike racks are statistically significant. Online methods can successfully reach large numbers for preference surveys, but introduce several problems: respondent geographic diversity and an ambiguous sampling frame.

Keywords: Visual Preference Survey, GIS, crowdsourcing,

Research support: Taubman College of Architecture and Urban Planning
1. Introduction

The main goal of this paper is to demonstrate an innovative method for conducting and analyzing data from urban visual preference surveys. Measuring public preferences about urban environments has been a perennial interest to planners. The ongoing revitalization of urban neighborhoods, and the complete streets movement means many cities are currently considering the design of their streets. New methods are needed to better gauge public preferences in ways which involve larger numbers of citizens, and can draw more nuanced conclusions from their preferences. Previous efforts to measure public preferences have taken the form of visual preference surveys (A. Nelessen Associates Inc., 1993). In these cases, planners compile photos that are presented to viewers, who generally are asked to provide a rating on a Lickert scale. As described below, these techniques have been criticized in several ways.

The method demonstrated has several innovative features.

First, responses are solicited through a public website using crowdsourcing methods. This results in a large number of respondents, in this case some 7,447 people. Although individual-level demographic data was not collected from respondents, we will argue the extensive research demonstrating that different types of people hold largely similar environmental preferences mitigates this drawback somewhat. In addition, some context information is available, since website analytics derived from visitors’ behavior provide aggregate information about visitors’ physical location, visit duration, and referring websites.

Second, the images visitors voted on were drawn from Google Street View’s database, reducing the potential bias of image framing and selection effects (Crisman, 2006). Street View photos are taken from a fixed height and location near the street centerline, and the voting website allowed viewers to rotate through a 360-degree panorama (see Figure 1).

Third, a ranking algorithm was used to convert the voting data into a preference score. Unlike in many visual preference surveys where respondents rate each photo according to a scale, in this survey they were asked to pick between two randomly selected streets, producing a dataset of pairwise votes. These votes were processed through a ranking algorithm used widely to produce rankings from pairwise competitions (such as chess players), eliminating concerns about inter-rated scale inconsistency.

Finally, municipal geographic information system (GIS) data was used to generate a database of street attributes, which were related to the preference ratings through the regression
analysis. This links the previous separate fields of visual preferences and indicators for the built environment. This was sufficient to generate many indicators previous research has suggested are most correlated with preference ratings, while avoiding the need for street observation or human attribution to create these indicators. However, both the use of the Street View images and GIS data may introduce new sources of bias.

Together, these methodological innovations demonstrate how public and private spatial data infrastructures and greater interoperability are opening up new sources of information and new professional tools like mapping mashups for professional planners (Ferreira, 2008). The remainder of the introduction will briefly describe previous research in the two areas this paper builds on: visual preferences and GIS indicators of urban form. Previous research on preferences for streets and urban environments finds preferences are generally consistent across demographic categories. In addition, the built environment measures are derived from previous efforts to derive urban design characteristics from GIS data.

Research on visual preferences suggests they are generally consistent across the population, but that some preference relationships may be curvilinear. In the field of visual
preferences, a meta-analysis of 107 studies found a correlation between all demographic groups of .82 (Stamps III, 1999). The only sub-groups with a low correlation was designers when asked to evaluate avant-garde architecture. (-0.46). Other groups which differed were children under 12 years old and special interest groups. Research on crowding on Montreal streets found an inverted u-shaped relationship between density and desirability, and augmented photos with video clips. Although not looking at the same characteristics as this survey, respondent “gender, cultural background, and location did not have significant effect on expressed desirability” (Nouri, 2011).

A related stream of research has sought to not only quantify preferences, but also relate an overall score to quantified components of an urban scene. Forsyth et al. and Stangl observe urban design can be viewed from multiple perspectives, such as audits, assessments, workshops, and mapping (Forsyth, Jacobson, & Thering, 2010; Stangl, 2008). Even among quantitative measurements of urban form, diverse measurements are used at different scales and by different fields (Clifton, Ewing, Knaap, & Song, 2008).

This paper follows a researchers who have sought to quantify preferences using variables derived from urban design theory, in particular the concepts of imageability, enclosure, human scale, transparency, and complexity. Ewing has decomposed ratings into environmental factors like this study, but differed in that static images and video clips were chosen and researchers manually coded street attributes (R. Ewing, King, Raudenbush, & Clemente, 2005). The study also applied a multilevel statistical model. Since individual-level votes were not recorded in this case, a hierarchical model could not be applied. However, the use of pairwise votes eliminates the bias of inter-rater scale differences, and the large number of voters and votes reduces the bias of unusual viewers. This study identified the following significant variables in explaining street preferences: proportion of street frontage occupied by parked cars, proportion of street frontage covered by tree canopy, curb extensions visible, proportion of buildings that house commercial uses, average sidewalk width, number of travel lanes, proportion of street frontage made up of dead spaces, underground utilities, quality of pavement maintenance. Working with GIS data in New York City, a research group demonstrated that many of the indicators measuring physical features predicting urban design ratings in Ewing et al. (2005) could be replicated with GIS data (Purciel et al., 2009).
Finally, this study builds on other scholarly works which have used Google Street View as a novel data source (Badland, Opit, Witten, Kearns, & Mavoa, 2010; Odgers, Caspi, Bates, Sampson, & Moffitt, 2012; Rundle, Bader, Richards, Neckerman, & Teitler, 2011). Researchers have found Google Street View can be used to reliability conduct expert audits of the built environment (Kelly, Wilson, Baker, Miller, & Schootman, 2013), or through crowdsourcing reliable identify accessibility barriers (Hara, Le, & Froehlich, 2013).

2. Methods

This section describes the project’s research methods: selecting streets, voters, analyzing preference data, calculating GIS urban design measures, and relating preferences to urban design.

2.1. Streets

The street voting data was collected via the “Beautiful Streets” website created by OpenPlans, a New York-based nonprofit (See Figure 2). OpenPlans wrote a website using the Google Maps API to query the Street View image for each of the points, recording the votes in a SQL database. The locations surveyed was created by taking a random sample of 200 street segments within Philadelphia city limits. Longer segments were subdivided into multiple points, which results in a total of 362 observation points (see Appendix B). Other visitor characteristics were recorded through tracking code associated with Google Analytics. Since the points were drawn exclusively from Philadelphia, the dataset contains less variation on several built environment variables than either the entire metropolitan area, or other cities with higher densities or different street designs. As discussed below, new tools mean this study could be easily replicated for additional places.
2.2. Visitors

Visitors were not required to register before voting, therefore no individual demographic data was collected. However, statistics about visitors’ physical location, the referring webpage, and other characteristics can be collected using a embedded tracking code through the Google Analytics service. According to this data, 63% of visitors were from the United States and 37% from other countries (see Table 1). The states with the most traffic were New York (where OpenPlans is based), Pennsylvania, and California. Most non-US visitors were from Canada or the United Kingdom. Visitors heard about the website through a variety of means. Twenty two percent were “direct visitors” who either typed the URL or the origin could not be determined. Thirty two percent were referred by popular social media and search websites. Thirty five percent were referred by nine popular urban-focused online publications that published articles containing links to the website. These websites have readerships interested in architecture, urbanism, and urban planning. If these readers have unrepresentative views, the results of the entire survey may be skewed. However, as discussed above, previous research suggests preferences are consistent across many socioeconomic variables. Designers have divergent views on innovative design, but the street database contains few streets with unusual designs.
Table 1. Geographic locations of survey website visitors. Source: Google Analytics.

<table>
<thead>
<tr>
<th>Country/Territory</th>
<th>Visits</th>
<th>Percent of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>4,704</td>
<td>63%</td>
</tr>
<tr>
<td>- New York</td>
<td>874</td>
<td>12%</td>
</tr>
<tr>
<td>- Pennsylvania</td>
<td>828</td>
<td>11%</td>
</tr>
<tr>
<td>- California</td>
<td>452</td>
<td>6%</td>
</tr>
<tr>
<td>- Massachusetts</td>
<td>216</td>
<td>3%</td>
</tr>
<tr>
<td>- Texas</td>
<td>215</td>
<td>3%</td>
</tr>
<tr>
<td>- Georgia</td>
<td>211</td>
<td>3%</td>
</tr>
<tr>
<td>- DC</td>
<td>197</td>
<td>3%</td>
</tr>
<tr>
<td>- Illinois</td>
<td>161</td>
<td>2%</td>
</tr>
<tr>
<td>- Washington</td>
<td>127</td>
<td>2%</td>
</tr>
<tr>
<td>- Minnesota</td>
<td>122</td>
<td>2%</td>
</tr>
<tr>
<td>- Virginia</td>
<td>119</td>
<td>2%</td>
</tr>
<tr>
<td>- New Jersey</td>
<td>107</td>
<td>1%</td>
</tr>
<tr>
<td>- Maryland</td>
<td>102</td>
<td>1%</td>
</tr>
<tr>
<td>- All Other US States</td>
<td>973</td>
<td>13%</td>
</tr>
<tr>
<td>Other Countries</td>
<td>2,743</td>
<td>37%</td>
</tr>
<tr>
<td>Canada</td>
<td>1,032</td>
<td>14%</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>350</td>
<td>5%</td>
</tr>
<tr>
<td>France</td>
<td>143</td>
<td>2%</td>
</tr>
<tr>
<td>Australia</td>
<td>131</td>
<td>2%</td>
</tr>
<tr>
<td>Brazil</td>
<td>89</td>
<td>1%</td>
</tr>
<tr>
<td>Germany</td>
<td>88</td>
<td>1%</td>
</tr>
<tr>
<td>Israel</td>
<td>82</td>
<td>1%</td>
</tr>
<tr>
<td>Netherlands</td>
<td>71</td>
<td>1%</td>
</tr>
<tr>
<td>Spain</td>
<td>69</td>
<td>1%</td>
</tr>
<tr>
<td>All Other Countries</td>
<td>688</td>
<td>9%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>7,447</td>
<td></td>
</tr>
</tbody>
</table>

Most website traffic, and presumably votes, were collected during the first two weeks after the website was launched. Data about the number and frequency of votes suggest most visitors only stayed long enough to contribute a few votes, however a small group of visitors voted many times (see Table 2). Eighty nine percent of visitors visited only once, and 87 percent stayed for less than 30 seconds. The frequency distribution of both shows steep negative exponential distributions. Given the website's novelty, it is expected that these visitors contribute one or two votes each. About three percent of visitors remained longer than 10 minutes.
<table>
<thead>
<tr>
<th>Visit Duration</th>
<th>Visits</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-10 seconds</td>
<td>6,190</td>
<td>83%</td>
</tr>
<tr>
<td>11-30 seconds</td>
<td>322</td>
<td>4%</td>
</tr>
<tr>
<td>31-60 seconds</td>
<td>160</td>
<td>2%</td>
</tr>
<tr>
<td>1-3 minutes</td>
<td>284</td>
<td>4%</td>
</tr>
<tr>
<td>3-10 minutes</td>
<td>253</td>
<td>3%</td>
</tr>
<tr>
<td>10-30 minutes</td>
<td>149</td>
<td>2%</td>
</tr>
<tr>
<td>More than 30 minutes</td>
<td>89</td>
<td>1%</td>
</tr>
</tbody>
</table>

Table 2. Survey website visitor duration. Source: Google Analytics.

Voting data was extracted from the website at three points, allowing a comparison between the numbers of cumulative visitors and voting data. Within three days of the official launch on 14 February, 2012, the website had already recorded 44% of the total votes. By 3 March, another 51% of votes were recorded, and the remaining 5% between 3 March and 18 May. The average number of votes per visitor, computed from these totals, decreased over this period from 20, to 15, and then to 3.

2.3. Vote Data

Visitors to the website voted on 103,000 pairs of Street View images. Two beauty indices were created from this data for comparison purposes. First, the percentage of appearances that resulted in votes was calculated. Second, the votes were processed using the Elo scoring algorithm. The Elo algorithm was invented by physicist and chess player Arpad Elo in order to rank chess players, and is suited for producing ranked orders from pairwise competitions. The Elo is implemented by assigning each point an equivalent score. The probability of success is computed based on each observation point’s previous score. Based on the result, the winner receives additional points that are subtracted from the loser. The total number of points shifted between the two is limited to 32 (the “k-factor”). The algorithm results in small changes if the outcome is similar to the predicted outcome, but more points are allocated if the outcome is unlikely based on the observation point’s existing scores. The second beauty index calculated was a simple average of the total number of favorable votes out of the total number of “appearances.” The Elo score and average number of votes have a Pearson’s Correlation coefficient of 0.9681.
The ELO score is computed as follows. Each site is assigned an initial score of 1200, which is updated after each pairwise vote. The ranking score of site A is updated to $R_A'$ according to the following formula after a vote against site B:

$$R_A' = R_A + K(S_A - E_A)$$

Where

$S_A$ is the actual outcome (either 0 or 1) and $E_A$ is the expected outcome, calculated as:

$$E_A = \frac{Q_A}{Q_A + Q_B}$$

And

$$Q_A = 10^{\frac{R_A}{400}}$$

$$Q_B = 10^{\frac{R_B}{400}}$$

The sites that result from this analysis seem to confirm the variable selection. Shown in Figure 3, the top two highest scoring sites are a site on Rittenhouse Square, and on a narrow downtown street, and the lowest scoring street was on the Schuylkill Expressway. The features visible in both high-scoring observation sites include street trees, a well-defined street wall, and in the case of Rittenhouse Square, bike parking and an active use (sidewalk café).
2.4. Urban Design Indicators

The urban design indicators constructed for this paper draw in particular on two previous studies of streets and urban design (Reid Ewing & Handy, 2009; R. Ewing et al., 2005). In Ewing et al. (2005), the authors measured features of main streets and their immediate environment to explain the main street scores rated by the 59 viewers of provided street photographs and video clips. Ewing & Handy (2009) developed an observation protocol, a list of urban design variables, to measure the five urban design constructs -- imageability, enclosure, human scale, transparency, complexity. We draw spatial indicators from both of the studies because they complement each other in a sense that best apply to our study: the observation
protocol in Ewing & Handy (2009) includes a more comprehensive list of variables to measure the urban street environment but was limited to commercial streets, and the variables in Ewing et al. (2005) focus on the immediate environment of streets but the object of study varies across all kinds of streets.

In order to construct GIS measures, these variables were operationalized with methods adopted from Purciel et al. (2009), a study that validated the consistency of using GIS measures as a proxy of Ewing’s observational measures of urban design. We compute most of the variables by either repeating Purciel et al.’s methods or with some modifications. Some variables such as number of people and noise level estimate are not computed due to their high complexity and data limitations. We also compute some of the variables that were not done by Purciel et al. by using proximate measurements that utilize existing datasets. To compute the variables, a 200-foot buffer was constructed around each observation point, an estimate of the probable visual range from the Street View panorama. Spatial data was obtained from the Pennsylvania Spatial Data Access (PASDA), a clearinghouse of spatial data operated by Pennsylvania State University, the City of Philadelphia, and a public street tree database. The spatial indicators constructed are summarized in Appendix A.

3. Results

This section reports the results of the regression analysis of the Elo voting scores and the urban design indicators described above. Three models were fit with these variables (See Table 3). Model 1 includes 21 variables and has an adjusted $R^2$ of 0.403. Model 2 omits four variables which could not be computed for every observation point, and has an adjusted $R^2$ of 0.440 and includes all 362 observation points. Finally, Model 3 includes only the variables from Model 2 whose coefficients were significant at the 95% confidence level, and the adjusted $R^2$ remains 0.440.
In Model 3, ten of the variables are significant, and all but one of the significant variables have the expected sign. The variables with the largest standardized coefficients (greater than 0.20 in absolute value) are number of trees, number of intersections, total street length (negative), proportion of properties with windows (negative), and proportion of built land. Appendix C contains a map illustrating tree and intersection density. The only variable without the expected sign is proportion of properties with windows. Only two of the variables in Model 3 have a pairwise correlation coefficient greater than 0.5: proportion of street segment with active uses

### Table 3. Summary of regression analysis, dependent variable the Elo scores, sorted by absolute value of Beta coefficients in Model 3.

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>310</td>
<td>362</td>
</tr>
<tr>
<td>Adj. R²</td>
<td>0.403</td>
<td>0.440</td>
</tr>
<tr>
<td>B</td>
<td>Beta</td>
<td>B</td>
</tr>
<tr>
<td>Number of trees</td>
<td>11.65***</td>
<td>0.363</td>
</tr>
<tr>
<td>Number of intersections</td>
<td>20.32*</td>
<td>0.204</td>
</tr>
<tr>
<td>Total street length</td>
<td>-0.09</td>
<td>-0.182</td>
</tr>
<tr>
<td>Proportion of properties with windows</td>
<td>-577.88***</td>
<td>-0.247</td>
</tr>
<tr>
<td>Proportion of built land</td>
<td>247.35**</td>
<td>0.133</td>
</tr>
<tr>
<td>Percentage of park and open space land</td>
<td>275.26**</td>
<td>0.120</td>
</tr>
<tr>
<td>Number of historic buildings</td>
<td>8.34***</td>
<td>0.216</td>
</tr>
<tr>
<td>Proportion of properties with active uses</td>
<td>135.03*</td>
<td>0.106</td>
</tr>
<tr>
<td>Number of buildings</td>
<td>4.52**</td>
<td>0.127</td>
</tr>
<tr>
<td>Number of bike racks</td>
<td>9.90</td>
<td>0.089</td>
</tr>
<tr>
<td>Proportion of street segments with bike lanes</td>
<td>37.26</td>
<td>0.031</td>
</tr>
<tr>
<td>Number of buildings with identifiers</td>
<td>-0.70</td>
<td>-0.041</td>
</tr>
<tr>
<td>Average height of buildings weighted by building area</td>
<td>0.182</td>
<td>0.045</td>
</tr>
<tr>
<td>Proportion of street segments with side parking</td>
<td>95.73</td>
<td>0.058</td>
</tr>
<tr>
<td>Visibility of major landscape feature within 551 feet</td>
<td>-29.53</td>
<td>-0.031</td>
</tr>
<tr>
<td>Number of buildings with non-rectangular</td>
<td>-0.88</td>
<td>-0.022</td>
</tr>
<tr>
<td>Proportion of dead spaces</td>
<td>-331.80*</td>
<td>-0.100</td>
</tr>
<tr>
<td>Diversity of building materials</td>
<td>-143.98**</td>
<td>-0.124</td>
</tr>
<tr>
<td>Diversity of building age</td>
<td>-1.08</td>
<td>-0.076</td>
</tr>
<tr>
<td>Quality grade related to building workmanship and maintenance</td>
<td>-4.64</td>
<td>-0.008</td>
</tr>
<tr>
<td>Street width weighted by street segments length</td>
<td>0.34</td>
<td>0.038</td>
</tr>
</tbody>
</table>

*** Significant at 99% (t > 2.576), ** Significant at 95% (t > 1.960), * Significant at 90% (t > 1.645).
and proportion of properties with windows have a correlation coefficient of 0.567, and number of intersections and total street length have a coefficient of 0.849. Therefore the coefficients of these variables may be unreliable, but multicollinearity will not effect the explanatory power of the entire model.

4. Discussion

This section will discuss several aspects of the results: the limitations of the analysis, the results of the regression analysis, and the replicability of this type of analysis.

4.1 Limitations

Our methodology differs from that of Purciel et al.’s in that they use the block face as the study unit while our measurements are based on a 200-foot buffer area from the observation points. As explained above, the logic behind this is that 200 feet is an estimate of the probable visual range from the Street View panorama. Because of this critical difference, we have made some adjustments to the variables measured by Purciel et al. and to their corresponding computing methods. For example, for the variable “number of courtyards, plazas and parks on the block face,” we measure “proportion of parks and open space land” instead because we feel it makes better sense in our case. Another typical example is that we use “proportion of properties with windows” to replace “proportion of street segments with windows.”

However, there are some inherent problems with this methodology. First of all, we cannot tell if all raters had rotated the Street View when they were voting for the photos. If they did not rotate the Street View, which means that they were only comparing at most half of (180 degree) the buffer areas, it inevitably impairs the validity of our assumption -- that the urban environment within the 200-foot buffer area could explain what the raters saw. Unfortunately, this inherent issue also forces us, in a sense, to use the 200-foot buffer area as our study area because we do not have the information regarding to which angles of the Street View photos the raters were looking at. Another related issue is that summarizing all urban environment objects within a 200-foot buffer area includes some objects that were not visible from the observation points. For example, some objects that are blocked by large buildings from the observation points will still be counted into our measurements if they are within the buffer. This problem becomes extremely serious when they are many objects away from the observation points and blocked by buildings.
We suspect that the variable “proportion of properties with windows” is negatively associated with the ELO score is exactly because of this problem. Theoretically, this issue could be dealt with by running a viewshed analysis for each of the observation points, and excluding features that were not visible.

4.2 Results

This section will briefly consider the explanatory power and validity of the regression analysis, and then consider the substantive implications regarding urban street preferences. First, the overall explanatory power of the model (Adj. $R^2$ of 0.440) suggests many relevant factors have been included, it might be improved in several ways. First, although scatterplots did not suggest it was needed, Nouri (2011) suggests some preferences may take curvilinear forms, such as the inverted “U” that study observed for pedestrian density. Additional variables which could not be computed here which would likely add explanatory power include the presence of pedestrians, sidewalk cafes, and greater details about sidewalk width and bump-outs. In addition, a more precise measurement of the street wall might improve the explanatory power of that variable. Finally, subsequent research could explore threshold effects or interaction effects among variables. It could be that groups of variables might together explain more than they each do individually. This could especially be true if observers are providing ratings by how well the scenes match ideal types they hold in their minds (see below).

Considering the variables and coefficients estimated here suggests a variety of interesting interpretations. Construed as a measure of urban form, this study finds that the voting data revealed forms advocated by advocates for New Urbanism and traditional neighborhood design: small or nonexistent building setbacks to create enclosed streets, plentiful street trees, and historic buildings with traditional architecture as well as parks and open spaces. In addition, the data may suggest that bicycle infrastructure can contribute to perceived beauty – the variable for bicycle racks was significant, and although the presence of bike lanes was not significant at 90% it had a relatively high t-score in Model 2 (1.568). These general results could be subject to three general interpretations:

First, the respondents to this survey may have more “urban” preferences than the public at large. Since the website was created and promoted through websites and publications focusing on urban design, planning, and policy, there could be a self-selection bias among the
respondents. We are not aware of any systematic studies of professional attitudes, and question the assumption in the more general literature on environmental preferences that most people are similar. Preferences may have shifted in the ensuing decades from many of those research, and the original studies are generally limited to small groups. The method demonstrated here makes it possible to conduct this research with much larger numbers of people.

Second, there may be a separation between street preferences and neighborhood preferences. Neighborhoods with beautiful streets may be preferred for shopping, work, or recreation, but not as a residential context. Many lively commercial districts have traditional urban form, such as Georgetown in Washington, D.C., Manhattan, and even Disney World’s “main street,” however preferences for residential neighborhoods may be different.

Third, this data could be documenting a gap in preferences between the public and the available urban spaces, which might help explain the empirical presence of a walkability premium documented in the literature (e.g., Pivo & Fisher, 2011). However the literature on residential choice also indicates relevant factors for neighborhood preferences include non-aesthetic factors such as school quality, tax rates, crime rates, etc. Studies which integrate aesthetic and non-aesthetic considerations could shed light on the details and relative magnitude of both factors.

Finally, this paper does not engage the more theoretical literature on environmental preferences, although the method demonstrated here could be adapted to bring fresh perspective on the issue. An older body of research proposed a universal behavior-response model based in an evolutionary perspective proposing a “type of rapid, unconscious type of cognition” preceded judgments of the environment (Kaplan, 1987). Subsequent work informed by urban design theory, especially the work of Kevin Lynch, stressed the importance of cognition for understanding responses to particular urban places (Nasar, 1998). However none to our knowledge address the uniquely value-, and experience-laden nature of urban environmental preferences.

4.4 Replicability of the Method

This section considers how easily the methodology described here could be replicated, whose novelty lies in combining several techniques to arrive at a conclusion. First, although the original website used to create this website is no longer available, the technology to conduct
these surveys is available. Jennifer Evans-Cowley at Ohio State University has created a tool called StreetSeen (http://streetseen.osu.edu), which allows visitors to conduct their own surveys using Street View images for any area of their choosing. The software which powers this website has been published under an open license, making it available for others to build on. A research group at the MIT Media Lab has also created a similar website called PlacePulse, where they ask visitors to rate photos from 56 cities according to five dimensions: which looks safer, wealthier, more boring, livelier, and more depressing. Although the source code of their tool has not been released, they have released data from an earlier pilot study.

Computing the urban design indicators will introduce additional challenges. Although many of the required layers were freely available, several require detailed property data which required payment of $100 and a lengthy delay for the data to be delivered through the mail. Other data which was difficult to obtain was street speed limits, widths, and the presence of extensions. We were unable to obtain a list of restaurants with sidewalk café licenses despite multiple requests, which prevented the calculation of a variable with outdoor dining.

Furthermore, the analysis results may not take the form most useful from a planning perspective. After hearing an earlier version of this paper, a professional audience suggested a useful feature would be to test changes in preferences to urban design changes, such as street improvements, benches, redesigned streets, and trash cans. However, achieving this would require simulated images, or collecting before-and-after Street View photos, a feature now not available from Google. Professionals also stated a desire to estimate willingness to pay, to meet the perceived demand by public officials for financial perspectives on public improvements.

4.4 Towards Data Infrastructures for Planning

Finally, this project relies on a growing data infrastructure for planning: municipal and statewide GIS programs which distribute spatial information, private sector data providers such as Google, and the creators of mashup websites that combine data from multiple sources to create tools for professional goals, such as here conducting a survey. Although the data and methodologies are well known, this type of survey is not yet widespread in practice. Visual preference surveys play only one small part of a planning process, and may only be useful for projects which seek to address aesthetic considerations.
The demonstration project reported here also illustrates one problem with crowdsourced data collection methods: the ambiguous and often widely distributed geographic location of participants. This paper therefore illustrates an emerging issue in professional practice: how online participation opportunities has challenged assumptions that participants live in the or near the area being planned. This mismatch between what Melvin Webber called the “community of propinquity,” or local geographic communities which are generally the site of planning activities, and the online “nonplace urban realm” of interested citizens (Webber, 1964). Given the high levels of mobility in society, this means online tools can potentially help reach geographically dispersed stakeholders, who might include in addition to full-time residents people who work of visit in an area, part-time residents, or those with sentimental attachment (such as former residents). To our knowledge this issue has yet to be adequately explored by researchers in either empirical or theoretical terms.

5. Conclusions

This paper describes a novel method for conducting a crowdsourced visual preference survey. Participants are asked to vote on a public website among pairs of Google Street View panoramas for a collection of randomly selected streets in Philadelphia. The resulting votes are processed using an Elo ranking algorithm to produce a preference score. Following Ewing et al. (2005), these scores are related to urban design indicators computed from GIS data.

The statistical analysis finds that roughly 44% of the variation in preference scores can be explained by ten variables: the number of trees, number of intersections, street length, proportion of properties with windows, proportion of built land, percentage of park and open space, number of historic buildings, number of properties with active uses, number of buildings, and number of bike racks. The explanatory power of the model could be improved through better and additional indicators, however urban visual preferences may retain a holistic character difficult for quantitative research such as this to fully capture.

Professional visual preference surveys, and a large body of environmental preference research, rests on experiments with relatively small numbers of participants. The method demonstrated here illustrates how analysis can be used to make sense of data that results from Street View preference surveys. It calls for the further development of tools to explore visual preferences for research and professional purposes.
Acknowledgements

The authors wish to thank Tharron Combs for assistance in the literature review, Andy Mehos and Brian Ivey for assistance in obtaining needed GIS information, and Aaron Ogle and Mjumbe Poe, the original creators of the Beautiful Streets website, for providing the voting data for the analysis, and answering questions about their tool. We are responsible for any errors in the analysis.


## Appendix A

Summary of variable data sources, methods, mean, and standard deviation.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Data Source(s)</th>
<th>Method</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELO beauty score</td>
<td>See text</td>
<td>See text</td>
<td>1200.27</td>
<td>252.70</td>
</tr>
<tr>
<td>Percentage of park and open space land</td>
<td>PASDA Philadelphia land use dataset</td>
<td>Export land use type of &quot;park and open space&quot; from land use dataset and intersect it with 200-feet buffer, dissolve by observation point ID</td>
<td>0.046</td>
<td>0.131</td>
</tr>
<tr>
<td>Visibility of major landscape feature within 551 feet</td>
<td>PASDA Philadelphia hydrography line and poly dataset</td>
<td>Create a 551-feet (168m) buffer from the observation points, intersect with the hydrography dataset, dissolve by observation point ID</td>
<td>0.097</td>
<td>0.296</td>
</tr>
<tr>
<td>Proportion of built land</td>
<td>PASDA Philadelphia Buildings dataset</td>
<td>Summarize the total area of building footprint and divided it by the total area of 200-buffer area</td>
<td>0.222</td>
<td>0.139</td>
</tr>
<tr>
<td>Proportion of properties with windows</td>
<td>PASDA Philadelphia Parcel dataset, Philadelphia Office of Property Assessment (OPA) Certified Property Assessment 2014 dataset - &quot;Property_type&quot; attribute</td>
<td>Join Parcel dataset with Property Assessment table, count the number of properties that are in the type of &quot;Commercial, restaurants, stores, and offices,&quot; and divide the number by the total number of properties</td>
<td>0.042</td>
<td>0.102</td>
</tr>
<tr>
<td>Proportion of properties with active uses</td>
<td>PASDA Philadelphia Parcel dataset, Philadelphia OPA Certified Property Assessment 2014 dataset - &quot;Property_type&quot; attribute</td>
<td>Join Parcel dataset with Property Assessment table, count the number of properties that are in the type of &quot;Apartments, Bank &amp; Office Buildings, Miscellaneous, commercial, Condominium, Health Facilities, restaurants, schools, public utilities, stores, and offices,&quot; and divide the number by the total number of properties</td>
<td>0.115</td>
<td>0.183</td>
</tr>
<tr>
<td>Proportion of street segments with side parking</td>
<td>Philadelphia Parking Authority Philadelphia Parking Meters Inventory June 2013</td>
<td>Intersect Street Centerline dataset with 200-feet buffer, join intersect result with Parking Meters table, then assign &quot;1&quot; for matched records and &quot;0&quot; for unmatched ones, summarize total length of street segments with value &quot;1&quot;, and finally divide it by total street segments length</td>
<td>0.047</td>
<td>0.143</td>
</tr>
<tr>
<td>Proportion of street segments with bike lanes</td>
<td>PASDA Philadelphia bike network dataset, PASDA Philadelphia street centerlines dataset</td>
<td>Merge &quot;PhiladelphiaBikeTrailsSidePaths&quot; and &quot;PhiladelphiaBikeNetwork&quot; layers in Bike Network dataset, intersect the result with 200 buffer, summarize the total length of bike lanes in each point buffer, and divide it by total street segments length</td>
<td>0.076</td>
<td>0.195</td>
</tr>
<tr>
<td>Proportion of dead spaces</td>
<td>PASDA Philadelphia land use dataset</td>
<td>Export land use type of &quot;Vacant&quot; from Land Use dataset, intersect it with 200-feet buffer, then dissolve the result by observation point ID, and finally divide the summed result by 200-feet buffer area</td>
<td>0.041</td>
<td>0.073</td>
</tr>
<tr>
<td>Number of buildings</td>
<td>PASDA Philadelphia Buildings dataset</td>
<td>Intersect Buildings dataset with 200-feet buffer, and dissolve by observation point ID</td>
<td>9.572</td>
<td>7.019</td>
</tr>
<tr>
<td>Topic</td>
<td>Description</td>
<td>Method</td>
<td>Result 1</td>
<td>Result 2</td>
</tr>
<tr>
<td>-----------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>Quality grade related to building workmanship and maintenance</td>
<td>PASDA Philadelphia Parcel dataset, OPA Expanded 2014 Property Assessment dataset - &quot;Exterior Condition&quot; attribute. Join the Parcel dataset with Property Assessment data, intersect it with 200-feet buffer, and calculate the mean quality grade of all properties that have values in their &quot;exterior condition&quot; field.</td>
<td></td>
<td>3.959</td>
<td>0.436</td>
</tr>
<tr>
<td>Diversity of building materials</td>
<td>PASDA Philadelphia Parcel dataset, OPA Expanded 2014 Property Assessment dataset - &quot;General Construction&quot; attribute. Join property assessment data with Parcel dataset, intersect Joined Parcel layer with 200-feet buffer, then export intersect result as a table, and finally in Microsoft Excel using Simpson Diversity Index to calculate diversity.</td>
<td></td>
<td>0.809</td>
<td>0.202</td>
</tr>
<tr>
<td>Diversity of building age</td>
<td>PASDA Philadelphia parcels dataset, OPA Expanded 2014 Property Assessment dataset - &quot;Year Built&quot; attribute. Join the Parcel dataset with Property Assessment data, intersect it with 200-feet buffer, and dissolve (Standard Deviation is used as the statistical type) the result by observation point ID.</td>
<td></td>
<td>14.706</td>
<td>16.457</td>
</tr>
<tr>
<td>Number of historic buildings</td>
<td>PASDA Philadelphia registered historic sites - Philadelphia register, National Park Service, National Register of Historic Places. Use spatial join to count the number of historic buildings from the two datasets.</td>
<td></td>
<td>1.489</td>
<td>5.706</td>
</tr>
<tr>
<td>Number of buildings with identifiers</td>
<td>PASDA Philadelphia Parcel dataset, Philadelphia Office of Property Assessment (OPA) Certified Property Assessment 2014 dataset - &quot;Property type&quot; attribute. Join Parcel dataset with Property Assessment table, and count the number of properties that are in the categories of &quot;Amusements, Automotive, Commercial, Houses of Worship, Office and/or Hotel, Miscellaneous, Restaurants, Stores, Stores and Offices, Stores/ office, hotel/ rooming house&quot;.</td>
<td></td>
<td>4.033</td>
<td>13.072</td>
</tr>
<tr>
<td>Number of buildings with non-rectangular shapes</td>
<td>PASDA Philadelphia Parcel dataset, OPA Expanded 2014 Property Assessment dataset - &quot;shape&quot; attribute. Join Parcel dataset with Property Assessment table, and count the number of properties whose shape are not rectangular.</td>
<td></td>
<td>4.851</td>
<td>5.809</td>
</tr>
<tr>
<td>Average height of buildings weighted by building area</td>
<td>PASDA Philadelphia Parcel dataset, OPA Expanded 2014 Property Assessment dataset - &quot;stories&quot; attribute. Join Parcel dataset with Property Assessment table, add a &quot;height&quot; field, the value of which is estimated by multiplying 12.14 feet (3.7m) with &quot;stories&quot; field, and the average building height is the average of buildings in the buffer, weighted by the proportion that each building's area represented of the 200-feet buffer.</td>
<td></td>
<td>52.708</td>
<td>65.403</td>
</tr>
<tr>
<td>Number of bike racks</td>
<td>PASDA Philadelphia bike racks dataset. Use 200-feet buffer to spatial join Bike Racks points layer.</td>
<td></td>
<td>0.599</td>
<td>2.262</td>
</tr>
<tr>
<td>Number of trees</td>
<td>Philadelphia Tree Map website. Use 200-feet buffer to spatial join Tree point layer.</td>
<td></td>
<td>7.669</td>
<td>7.376</td>
</tr>
<tr>
<td>Street width weighted by street segments length</td>
<td>PASDA Philadelphia street centerlines dataset, Philadelphia Department of Streets street width dataset. Join Street Centerlines dataset with Street Width table, intersect it with 200-feet buffer, and finally calculate the street width by weighting the that each street segments' length represented of the total length of all street segments.</td>
<td></td>
<td>35.742</td>
<td>12.218</td>
</tr>
<tr>
<td>Street speed limit weighted by street segments length</td>
<td>PASDA Philadelphia street centerlines dataset, Philadelphia Department of Streets street speed limit dataset. Join Street Centerlines dataset with Speed Limit table, intersect it with 200-feet buffer, and finally calculate the speed limit by weighting the that each street segments' length represented of the total length of all street segments.</td>
<td></td>
<td>19.250</td>
<td>12.218</td>
</tr>
<tr>
<td>Number of intersections</td>
<td>PASDA Philadelphia street centerlines dataset</td>
<td>Run &quot;Feature Vertices to Points&quot; tool with street centerline dataset, then run &quot;Collect Events&quot; with the output and delete events with an Icount&lt;2, use 200-feet buffer to spatial join (sum) the remaining points</td>
<td>2.337</td>
<td>2.332</td>
</tr>
<tr>
<td>-------------------------</td>
<td>---------------------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>Total street length</td>
<td>PASDA Philadelphia street centerlines dataset</td>
<td>Intersect Street Line dataset with 200-buffer, and dissolve by observation point ID</td>
<td>949.395</td>
<td>485.634</td>
</tr>
</tbody>
</table>
Appendix B.
Street View observation points.

Philadelphia
Visualizing a City Using Collection of Crowdsensed Sound

Itaru Usami  Niwat Thepvilojanapong*  Yoshito Tobe
Department of Integrated Information Technology
Aoyama Gakuin University, Japan
{itaru, yoshito-tobe}@rcl-aoyama.jp
*Mie University, Japan
wat@net.info.mie-u.ac.jp

Abstract
Urban environment is full of many kinds of sound. The sound can represent the atmosphere of a place; some place is quiet inside a park and another place may be very noisy with many passing cars. Thus collecting sound at many places help understand the atmosphere in the urban environment if the sound information is associated with time and place. Although several sounds maps exist, it is difficult to make a thorough map with reasonable density of information. Deployment of professional microphones increases cost. To solve the problem, participatory sensing or crowdsensing is expected to be useful. In the crowdsensing, people carrying mobile phones embedded with a microphone upload information at some certain place not only voluntarily but being asked with some tasks. As the number of people to join the sensing activity, a map will contain much information. Even with the mechanism of crowdsensing, we still have a problem of adding meaning to the sound if raw sound is only associated with a location. In this study we add meta-data to the sound to show the kind of the sound assuming that a sensing person uploads a set of information consisting of sound, location, and time. We aim to classify kinds of sound in order to make a sound map system that can be queried by users. The users are supposed to obtain attributes of a recorded location by
query. As the first step of the work, we have designed a method called iCSM to extract traffic of automobiles. The benefit of iCSM is to recognize the number of automobiles in a certain range. We are currently making a sound map focusing on the traffic of automobiles in our local city called Sagamihara.

Keywords:
crowdsensing, sound, classification

1. Introduction
Along with industrialization, city environments are rapidly changing. Once the number of car traffic has increased, it may be difficult to enjoy an atmosphere of forest. Thus the sound can represent the atmosphere of a place and collecting sound at many places help understand the atmosphere. This nature also has a regional aspect. For instance, we can hear much sound of motorbikes in the cities in Vietnam but do not hear much in other countries such as the Netherlands where commuting bicycles are more common. Thus collecting sound at many places help understand the atmosphere in the urban environment if the sound information is associated with time and place. A natural question is “why won’t we use photos instead of sound?” Our answer is “photos have much richer information than sound, but the privacy issue is difficult to overcome.”

For collecting sound at many places, we may need to deploy many professional microphones, which increases cost. Therefore, we utilize participatory sensing or crowdsensing (Campbell 2006) to increase the scalability of collection. In the crowdsensing, people carrying mobile phones embedded with a microphone upload
information at some certain place not only voluntarily but being asked with some tasks. As the number of people to join the sensing activity, a map will contain much information.

Based on the idea of using sound, there have been several activities of making sound maps. Sound maps provide many interesting features of recorded location, e.g., traffic of automobiles and walkers, population of living things and noise level.

Generally, sound map systems merely classify sound as where and when the sound are recorded and cannot classify features of recorded sound. Therefore, we aim to classify kinds of sound in order to make a sound map system that can be queried by users easily. The users are supposed to get easily attributes of recorded location by query to the system. As the first step of the work, we propose a method to extract traffic of automobiles. The benefit of this work is to recognize automobiles from traffic volumes.

In this study we add meta-data to the sound to show the kind of the sound assuming that a sensing person uploads a set of information consisting of sound, location, and time. We aim to classify kinds of sound in order to make a sound map system that can be queried by users. The users are supposed to obtain attributes of a recorded location by query. As the first step of the work, we have designed a method called iCSM to extract
traffic of automobiles. The benefit of iCSM is to recognize the number of automobiles in a certain range. Another important issue is the management of temporal property. Every place exhibits different types of sound depending on the time in a day or the day of the week etc. We create a simple data structure to manage the temporal axis. Another important aspect for this crowdsensing system is providing incentives to the participants (Yang 2012). We also focus on the incentives to consider spatial and temporal dynamics of participants.

We are currently making a sound map focusing on the traffic of automobiles in our local city called Sagamihara. The Web site is a part of Sagamihara vCityMap and Android phone application is available and anybody using an Android phone can upload sound data.

The rest of the paper is organized as follows. Section 2 explains works related to ours. Section 3 describes the design of iCSM. Sections 4 and 5 shows our current status of vCityMap and incentives.

2. Related Works
A work on sound classification has been done in Eronen, 2006. In the study, commonly-used significant metrics are calculated for various kinds of sound occurred in our daily lives and used as the inputs to Hidden Markov Model classifier. Unlike this work, our study focuses on sound in the outdoor environment for creating a sound map. Another area closely related to ours is participatory sensing or crowdsensing. Although the idea of utilizing citizens' activity with phones was shown in the past, Campbell
defined the problem well in Campbell 2006. Since then, there have been considerable amount of work in the area. Air quality was monitored with handheld mobile phones (Dutta 2009). The distribution of radio-frequency electromagnetic fields was investigated in the city Zurich (Haenfratz 2013). It also uses the regular operation of streetcars. Noisemap is being created to assess the environmental pollutants in a German city (Meurische 2013). Other studies also consider participants in vehicles. In Mohan 2008, a person inside a car is assumed to have a smartphone and an accelerometer inside the phone captures the conditions of road. The most closely related to ours is SoundSense (Lu 2009). We own our basic structure to this work. However, ours is different in that we show meta-data derived from raw sound and visualize the distribution in a map. In addition, we also define a query format to extract the characteristics related to places. Thus, there are many related works, but our study particularly focuses on query-able people-generated sound maps and not only on the noise level or the sound strength inside the city.

3. iCSM

Of all kinds of sound, we first consider extracting the number of passing cars because larger traffic indicates higher activities in the city. At the signal strength of sound, we assume that the volume of the recorded sound forms the shape of an arch when cars pass by a sound recorder or a microphone. Based on this feature, we propose a new method for accurately classifying passing cars as whether the volume of the recorded sound is forming the shape of an arch or not. The proposed algorithm is explained in the next subsection.
3.1 Preparation

We first extract a sound volume signal and apply Fourier transform to the originally collected signal. In the process of examining frequency spectra, we have found that car-related components are limited in the range of 2-3 kHz. The sound sources of cars consists of engine sound, wind noise and road noise. The frequency of wind noise is higher than 3 kHz. Therefore, we have decided to extract the 2-3 kHz components.

Fourier transform is applied to each window of 23 ms. Within each window, the mean value of intensity between 2-3 kHz is computed to extract a sound volume signal. We also apply first median filter, and then moving average filter to the subband mean time series and obtain a smoothed transition.

Next, we need to establish a method to detect the arch. First, we calculate the differentiation of the filtered signal. When the differentiation curve crosses zero line, it indicates either the top or bottom of the arch. Required arches are the period between two adjacent bottoms of the arch. However, small arches which do not result from the sound of cars, will also be detected by this method.

Let $I_i;i+1$ be the integral between the two adjacent zero-crossing points $i$ and $i+1$. Since some zero-crossing points are removed, $I_i;i+1$ and $I_i+1;i+2$ may be positive or negative. In this case, point $i + 1$ is removed. As a result, only large arches can be extracted. Next, we filter the candidate arches from the height and correlation of an arch. The height of arch is an integral of arches from the beginning of the arch to the end of the arch. The candidate arches are filtered if the height of arches is less than the threshold. After the filtering, non-arches may be included in the result. To remove such non-arches, we
calculate a correlation between the smoothed signal and the original signal. Therefore, most of spike noises should be removed and arches can be correctly extracted.

3.2 Classifier and Meta-Data Database

The calculated number of the arches corresponds to the number of passing cars. Therefore, we use the number as meta-data for the original sound. Finally, the original sound is associated with the location and time of acquisition and the meta-data, which will be stored in the "Geotagged Meta-Data DB".

3.3 Query Format

We assume that users easily retrieve beneficial information from sound maps. To design query-able database is essential for this purpose. For example, when executing a query of “where is a road with many cars”, locations where the condition is satisfied are extracted. We define a query syntax to retrieve data. The query is comprised of the following expressions.

- hFieldExpressioni : Automobile, Crowd, Animal
- hAreaExpressioni : World, Areas, Country
- hTimeExpressioni : Seasons, Morning, Noon, Night
- hQuantityExpressioni : Large, Small
- hAbstractionExpressioni : Noisy, Melodic

Field Expression specifies a kind of sound. Area and Time Expression restrict the retrieval range. Quantity Expression represents how many a kind of sound in Field Expression is included. For example, users can retrieve traffic volume and crowd volume by means of this expression. Abstraction Expression selects sound matching
specified abstraction. The query is used in combination of the above five expressions.

3.4 System Architecture

![iCSM Architecture Diagram]

**Figure 1** iCSM architecture

iCSM is a collection of the providers and visualizer of sound in an urban environment. Figure 1 shows the whole architecture. The server receives the uploaded raw sound signal and extracts the meta-data, which will be stored into the Geotagged Meta-Data DB. When a client issues a query such as "pick up areas with high traffic", the server
returns the matched areas and the corresponding map. The map in the server also provides raw sound data.

4. Sagamihara vCityMap

Sagamihara is a local city near Tokyo and the number of population is growing, which changes the city environment. iCSM helps understanding how the environment changes. We have created the above iCSM application software and delivered it on our Web page. For visualization, the OpenStreetMap (Hakley 2008) is being used and accessible by any citizen in Sagamihara. The service began in July 2014 as Sagamihara vCityMap and we need to recruit many participants from now on.

Figure 2  Sagamihara vCityMap
5. Incentives

As stated above, we need many participants to Sagamihara vCityMap. For that purpose, we are also developing a mechanism to encourage participation to the data collection in vCityMap. Many people are hindered from providing sensing data using smartphone due to battery consumption, time and monetary cost. Smartphone users may not participate in sensing tasks without payment of reward or remuneration in the real world.

A main approach for making participants get involved more actively is introducing an incentive mechanism which can encourage users to provide their data. In this project, we propose SenseUtil, a participation-aware incentive framework for crowd sensing.

The objectives of SenseUtil are increasing the number of sensing while keeping moderate payment. By applying the proposed mechanism, consumers who need data pay rewards to producers who carry out sensing tasks and report the data. SenseUtil applies the concept of microeconomics, where demand and supply determine the value of sensed data, e.g., if no one reports sensed data, the incentive is automatically increased in order to tempt inactive people. The incentive of SenseUtil, which dynamically changes with the time, depends on many factors including location, data types, elapsed time and users’ preference. Moreover, the incentive changes according to participation activity of producers because participation awareness is taken into consideration. In particular, SenseUtil determines the value of sensing activities by defining utility functions which are used to calculate economic reward. To study the impact of SenseUtil, we conducted a simulation study. The results show that historical data of participation activities help decrease payments moderately.

SenseUtil consists of three main players: consumers, producers and a server. A
A consumer would like to have data being sensed at a remote area, while a producer is willing to carry out such sensing tasks. A person can serve as both the consumer and producer. A central server is responsible to manage interactions between consumers and producers. Note that the notations of SenseUtil framework are given in Table 1. A consumer defines a Point of Interest (POI) where data should be sensed. In addition to location information, POI also includes starting time, expiry time and data type (i.e., which kind of data need to be sensed). The consumer sends POI information to the server on demand. When receiving corresponding data, the consumer pays a reward determined by the utility functions.

### Table 1: Summary of notation. (Each symbol is explained when it first appears in the paper.)

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_k$</td>
<td>Distance threshold of a producer $k$</td>
</tr>
<tr>
<td>$U_k$</td>
<td>Utility threshold of a producer $k$</td>
</tr>
<tr>
<td>$T_k$</td>
<td>Elapsed time threshold of a producer $k$</td>
</tr>
<tr>
<td>$U_i(t)$</td>
<td>Independent utility at time $t$</td>
</tr>
<tr>
<td>$t$</td>
<td>Current time</td>
</tr>
<tr>
<td>$t_{prev}$</td>
<td>Latest sensing time at POI $i$</td>
</tr>
<tr>
<td>$U_{min}$</td>
<td>Minimum utility</td>
</tr>
<tr>
<td>$U_{max}$</td>
<td>Maximum utility</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>A constant to determine increasing rate of utility</td>
</tr>
<tr>
<td>$U_{hist}(t)$</td>
<td>History-based utility at POI $i$ at time $t$</td>
</tr>
<tr>
<td>$U_{min}^{hist}$</td>
<td>Minimum utility for history-based utility</td>
</tr>
<tr>
<td>$U_{max}^{hist}$</td>
<td>Maximum utility for history-based utility</td>
</tr>
<tr>
<td>$U_{hist}$</td>
<td>Minimum utility in the record</td>
</tr>
<tr>
<td>$\delta_{min}, \delta_{max}$</td>
<td>Constants for adjusting minimum and maximum utilities</td>
</tr>
<tr>
<td>$D_{hist}$</td>
<td>Distance for calculating minimum incentive</td>
</tr>
<tr>
<td>$T_{hist}$</td>
<td>Period for calculating minimum incentive</td>
</tr>
</tbody>
</table>

A server is a middleman between consumers and producers. It maintains POIs’
information or the sensing tasks requested by consumers and updates corresponding reward of each POI periodically or on demand. The producers acquire detailed information of sensing tasks by exploiting pull and/or push services. By adopting the pull or on-demand services, the server dispatches POI information upon receiving a request from a consumer. The producers may use the pull service to avoid being overwhelmed by too frequent update of POI information.

In addition, the producers can use the pull service to update current reward of POI. On the other hand, the push service provides two methods for dispatching the information to producers, i.e., instant and periodic push. The instant push allows the server to dispatch the POI information immediately upon receiving new POI information from a consumer. The service is beneficial for producers who would like to have the information of new POI in real-time manner; thus they can act fast to receive rewards. The producers subscribe to periodic push will receive the POI information periodically. The server is responsible to collect sensing data from producers and forward the data to consumers. The process of collecting payment and rewarding are also handled by the server.

As described above, a producer receives the information of sensing tasks including current reward from the server through pull and/or push services. The producer can also determine her preferences including area of interest (e.g., a limited area based on current position or any specific area), maximum number of tasks, minimum reward, frequency of push-based notification, and so on. The behavior of a producer depends on current position and reward of the sensing task. A producer k carry out a sensing task if all the
following conditions are satisfied: (1) her position is not far from a POI, i.e., the distance between the producer and the POI is shorter than or equal to $D_k$, (2) the reward is higher than or equal to a threshold $U_k$, and (3) the time elapsed from previous sensing at the same POI is longer than $T_k$. The underlying reason of the third condition is to avoid too frequent sensing at the same POI. If the above conditions are satisfied, the producer changes planned original route by moving towards the POI, carries out the sensing task and moves towards the original destination. By default, the producer switches to the maximum moving speed in order to minimize moving time. However, the producer may move with the current speed if she is not in a hurry. After the task has been done, the producer receives reward via the server. Note that the producers may calculate utility by using Equation 1 introduced in Section 2.2. However, the producers may have incorrect values of utility because they do not know when other producers carry out the sensing tasks. The producers need to use the pull service to ask for current utility maintained by the server.

5.1 Independent Utility

Independent POI means sensing data at POI $i$ are independent of other POIs. Basically, independent utility of each independent POI is initialized to the minimum value ($U_{\text{min}}$) and increases with the time until reaching the maximum value ($U_{\text{max}}$). Equation 1 defines the utility of a POI at time $t$.

$$U(t) = \max(U_{\text{min}}, \min(U_{\text{max}}, a(t - t_{\text{prev}})))$$, (1)

where $t_{\text{prev}}$ is the latest sensing time at the POI and is initialized to the starting time of the POI. While the sensing task is not done, the utility increases with the time due to higher demand of consumers. A coefficient $a$ ($a > 0$), which is determined by the
consumer, determines how fast the utility increases. The consumer also decides $U_{\text{min}}$ and $U_{\text{max}}$ because the value of data sensed at each POI may be unequal. When a sensing task has been done, the utility is reset to the minimum value and starts to increase again. The underlying reason of Equation 1 is straightforward. Consumers would like to urge producers to carry out sensing tasks but they would like to avoid too frequent sensing which is not likely to give meaningful information for most of the applications. Because some kinds of sensing data do not change abruptly, it would be better to have an interval between each sensing. By applying the above equation, consumers pay less for each sensing if sensing interval is short while they pay more if the interval is long.

5.2 History-based Utility

The frequency of each sensing point differs and depends on many factors. Location of a sensing point is a main factor, i.e., a crowded area is likely to be sensed more frequent and vice versa. Paid incentive (Equation 1) is directly affected by the sensing frequency. Thus we decide to use the minimum of paid incentive in a nearby area to calculate the next round of utility as expressed in (Equation 2).

$$U_h(t) = \max(U_{\text{min}}^h, \min(U_{\text{max}}^h, a(t - t_{\text{prev}})))$$

where $U_{\text{min}}^h = U_{\text{hist}} - \Delta_{\text{min}}$, and $U_{\text{max}}^h = U_{\text{hist}} + \Delta_{\text{max}}$.  \hspace{1cm} (2)

$U_{\text{hist}}$ is the minimum of paid incentive within $D_{\text{hist}}$ meters of the POI in $T_{\text{hist}}$ seconds ago. $U_{\text{min}}$ and $U_{\text{max}}$ in Equation 1 is substituted with $U_{\text{min}}^h$ and $U_{\text{max}}^h$ which is calculated from $U_{\text{hist}}$ and $\Delta_{\text{min}}$ or $\Delta_{\text{max}}$. Based on the history-based utility, the utility of rarely sensed POIs is likely to start with high incentive while the utility of frequently sensed POIs is likely to start with low incentive. As a result, sensing
activities of producer are likely to be distributed in the area.

5.3 Economic Point System

Any kinds of currency including monetary currency, virtual currency and a point system can be applied to SenseUtil for payments and rewards. The point system is widely adopted by real-world stores and electronic commerce for a long time and has been proved to be a successful strategy to encourage purchases and maintain the loyalty of customers. If the point system is adopted, producers who receive points can use those points to request sensing data in the future.

6. Summary

We have proposed a system to show an urban environment with sound collected by crowdsensing. In particular, we focus on detecting the number of passing cars and therefore we have created a scheme to recognize the sound of car. For the first step towards making a sound map with distinguishable characteristics metrics, we have explored an algorithm to detect sound of passing-by cars. We have designed and implemented the algorithm as iCSM. It consists of Android applications in the smartphones and a server to analyze the data and manage the geotagged data. We have started a new service called Sagamihara vCityMap. In addition to the collection mechanism, we also create an effective incentive scheme to increase the number of crowdsensing participants.

References


CyberGIS-enabled Urban Sensing from Volunteered Citizen Participation Using Mobile Devices

Junjun Yin\textsuperscript{1,2,3, *}, Yizhao Gao\textsuperscript{1,2,3}, and Shaowen Wang\textsuperscript{1,2,3,4,5,6, *}

\textsuperscript{1}CyberGIS Center for Advanced Digital and Spatial Studies
\textsuperscript{2}CyberInfrastructure and Geospatial Information Laboratory
\textsuperscript{3}Department of Geography and Geographic Information Science
\textsuperscript{4}Department of Computer Science
\textsuperscript{5}Department of Urban and Regional Planning
\textsuperscript{6}National Center for Supercomputing Applications

University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

*To whom correspondence should be addressed: jyn@illinois.edu and shaowen@illinois.edu
Abstract

Environmental pollution has significant impact on citizens’ health and wellbeing in urban settings. While a variety of sensors have been integrated into today’s urban environments for measuring various pollution factors such as air quality and noise. To set up sensor networks or employ surveyors to collect urban pollution datasets remains to be costly and may involve legal implications. An alternative approach is based on the notion of volunteered citizens as sensors for collecting, updating and disseminating urban environmental measurements using mobile devices, which is also known as crowdsourcing. A big data scenario emerges, as large-scale crowdsourcing activities tend to generate sizable and unstructured datasets with near real-time updates. Conventional computational infrastructures are inadequate for handling such big data, for example, designing an one-fits-all database schema to accommodate diverse measurements, or dynamically generating pollution maps based on visual analytical work flows.

This paper describes a CyberGIS-enabled urban sensing framework to facilitate participation of volunteered citizens in monitoring environmental pollutions using mobile devices. Since CyberGIS is based on advanced cyberinfrastructure and characterized as high-performance, distributed, and collaborative GIS, the framework enables interactive visual analytics for big urban data. Specifically, this framework integrates a MongoDB cluster for data management (without requiring a predefined schema), a MapReduce approach to extracting and aggregating sensor measurements, and a scalable kernel smoothing algorithm using graphics processing unit (GPU) for pollution map generation. We demonstrate the functionality of this framework though a CyberGIS-enabled mobile application to a use case scenario for mapping noise levels by capturing geo-tagged and time-stamped noise level measurements as users move around in urban settings.

Keywords: Big spatial data, CyberGIS, environmental health, noise mapping, and urban sensing

1 Introduction

In today’s urban environments, various pollution problems have become significant concerns to people’s health and well-being. Being able to monitor and measure the status of environmental pollution with high spatiotemporal resolution for producing accurate and informative pollution maps is crucial for citizens and urban planners to effectively contribute to decision making for improving living quality of urban environments. Traditionally, government agencies are responsible for measuring and collecting urban pollution data, which is done either by employing surveyors with specialized equipment or by setting up monitoring networks. For ex-
ample, under the EU environmental noise directive (2002/49/EC) (Directive, 2002), some cities commenced the installation of permanent ambient sound-monitoring networks. This approach is, however, subject to several limitations. For example, it is often costly to build such sensor networks or hire many surveyors. Furthermore, such sensors are statically placed and each can only cover an area or space of certain size while sensor measurements are usually sampled and aggregated for a period of time resulting in low update frequency.

Due to these limitations, alternative approaches have been investigated including the utilization of citizens as sensors to contribute to collecting, updating and disseminating information of urban environments, also known as crowdsourcing (Howe, 2006; Goodchild, 2007). In particular, some previous studies have explored the idea of encouraging participatory noise monitoring using mobile devices. For example, NoiseTube mobile application utilizes the combination of microphone and embedded GPS receiver to monitor noise pollution at various sites of a city (Maisonneuve et al., 2009, 2010). This effort also showed some promising results regarding the effectiveness of participatory noise mapping. Compared to the traditional noise monitoring approach that relies on centralized sensor networks, the mobile approach is less costly; and with collective efforts, this approach using humans as sensors can potentially reach a significantly larger coverage of the city.

With integrated environmental sensors\(^1\), existing mobile devices (e.g. smartphones) can instrument comprehensive environmental properties, such as ambient temperature, air pressure, humidity, and sound pressure level (i.e., noise level). However, when a large number of participants are involved in crowdsourcing activities, a large volume of, near real-time updated, unstructured data sets are produced. Conventional end-to-end computational infrastructures will have difficulty coping with management, processing, and analysis of such data (Ball, 2012; Bryant, 2009), which requires support by CyberGIS geographic information systems and science based on advanced cyberinfrastructure (Wang, 2010; Wang, Anselin, et al., 2013).

This paper describes a CyberGIS-enabled urban sensing framework (Wang, Anselin, et al., 2013) to facilitate the participation of volunteered citizens in monitoring urban environmental pollution using mobile devices. This framework enables scalable data management, analysis, and visualization intended for massive spatial data collected by mobile devices. To demonstrate its functionality, this study focuses on the case of noise mapping. The framework integrates a MongoDB\(^2\) cluster for data storage, a MapReduce approach (Dean and Ghemawat, 2008) for extracting and aggregating noise records from mobile devices, and a parallel kernel smoothing algorithm using graphics processing unit (GPU) for creating noise pollution maps. The framework is implemented as a mobile application to capture geo-tagged and time-stamped noise level measurements as users move around in urban settings.

The remainder of this paper is organized as follows. Section 2 describes related work in the context of volunteered participation in sensing urban environment. We focus on research

\(^1\)http://developer.android.com/guide/topics/sensors/index.html

\(^2\)http://www.mongodb.org/
challenges in terms of data management, processing, analysis, and visualization. In particular, CyberGIS is argued to be suitable for addressing the challenges. Section 3 illustrates the details of the design and implementation of the CyberGIS-enabled urban sensing framework. Section 4 details a user case scenario for noise mapping using mobile devices. Section 5 concludes the paper and discusses future work.

2 Participatory Urban Sensing and CyberGIS

To monitor and study urban environmental pollution, data collection and processing are two important steps in our framework. In terms of data collection from volunteered citizens engaged in reporting noise levels around a city, researchers found a low cost solution of using the microphone of mobile device to record and calculate the sound levels, such as the SPL android applications\(^3\). Combining the embedded GPS receiver on mobile devices, the noise-level measurements are geo-tagged with geographic locations, which allow researchers to generate heatmap like noise maps (Maisonneuve et al., 2009; Stevens and DHondt, 2010).

With rapidly advancing mobile technologies, a variety of environmental sensors are equipped on today’s off-the-shelf mobile devices, which can measure ambient temperature, air pressure, humidity, and sound pressure level, etc. When a large number of volunteered citizens participate in sensing urban environments using mobile devices, it poses challenges for efficient data management. Because the availability of sensors varies in different devices, users measurements can seem to be unstructured, which makes it difficult to design an one-fits-all database schema to accommodate all the user inputs. A big data scenario emerges in large-scale crowdsourcing activities, which requires an innovative system to support scalable data handling, such as data storage with flexible data schema and efficient database querying. Many applications, such as NoiseTube, use relational database for data storage and processing, where relational databases with a rigidly defined, schema-based approach make it difficult to quickly incorporate new types of data (Stonebraker et al., 2007), and achieve dynamic scalability while maintaining the performance users demand (Han et al., 2011). The large volume and dynamic nature of the datasets also causes visualization problems for noise map generation. Existing GIS libraries, such as heatmap.js\(^4\) and map servers, such as GeoServer\(^5\) provide inadequate support for this type of data.

To embrace the characteristics of measurements from large-scale crowdsourcing activities and accommodate the geographic attributes of the user generated contents, CyberGIS integrates high performance computing resources and scalable computing architecture to support data intensive processing, analysis and visualization (Ghosh et al., 2012; Wang et al., 2012). CyberGIS represents a new-generation GIS based on the synthesis of advanced cyberinfrastruc-

\(^3\)http://play.google.com/store/apps/details?id=com.julian.apps.SPLMeter&hl=en
\(^4\)http://www.patrick-wied.at/static/heatmapjs/
\(^5\)http://geoserver.org/
A cutting edge CyberGIS architecture is shown in Figure 1 (Wang, Anselin, et al., 2013).

Our framework utilizes several components within this architecture. In the distributed data management component, we employ a MongoDB cluster for monitoring data intake and storage. Compared to relational database, NoSQL database supports more flexible data models with easy scale-out ability and high performance advantages (Han et al., 2011; Wang, Cao, et al., 2013). In the high performance support layer, we rely on MapReduce functionality of the MongoDB cluster for data processing, such as individual user trajectory extraction, which is used to visualize the pollution exposure to a particular participant; and data aggregation with an one-hour (this value is defined for the ease of implementation and can be changed according to user specifications) time window on the collected data provided by all participants, which is used to dynamically produce noise maps for the monitored environment. And finally, in the data and visualization layer, we apply a parallel kernel smoothing algorithm for rapid noise map generation using GPU. Specific design and implementation details will be discussed in the following section.

3 System Design and Implementation

The framework is designed and implemented to include two main components: a dedicated mobile application (for Android devices) for participants and a CyberGIS workflow for data management, processing and pollution map generation. A diagram for the overall architecture
is shown in Figure 2. We employ a service-oriented architecture for the integration between mobile devices and a CyberGIS platform.

Specifically, the mobile application utilizes the combination of GPS receivers and environmental sensors on mobile devices to produce geo-tagged and time-stamped environmental measurements. It is up to participants to decide when to upload their records to the CyberGIS platform via the implemented RESTful (Representational state transfer) web service interface. CyberGIS workflow filters and parses the input data (into JSON format) and store them into the MongoDB cluster. It also extracts trajectory of each individual participant to visualize the pollution exposure along the trajectory. For pollution map generation from the measurements that are uploaded by all of the participants, the data aggregation process is carried out using a specified time window. The pollution map is generated as a GeoTIFF image via a parallel kernel smoothing method using GPU, which will be displayed as a map overlay on top of the Esri world street map.

Figure 2: The architecture of the framework

3.1 CyberGIS Workflow

The workflow first filters out invalid data records (i.e., records without coordinates dropped) and then parses each record as a JSON object before saved to the MongoDB cluster. The MongoDB cluster is chained in a master-slave style in order to achieve scalability as datasets

---

6 http://json.org/
7 http://en.wikipedia.org/wiki/GeoTIFF
are accumulated into significant size. To visualize the pollution exposure to each individual user, we utilize MapReduce to simply use the device ID as the key to extract the trajectory of a specific user from the database. In this case, the reducer is optional. In relation to producing pollution maps for visualizing the current status of the measured environmental properties, we aggregate all users inputs based on a predefined time window and kernel bandwidth (e.g., sound decay distance). To simply the process, we define an one-hour time window and 50-meter kernel bandwidth. In other words, we assumed that each measurement will last for one hour and covers an area of 50-meter radius. The aggregation is implemented also using the MapReduce method, where the device ID is treated as the map key and the reduce process is based on the timestamps that fall in a specified one-hour time window.

The pollution map is dynamically generated by using a kernel smoothing method. Kernel smoothing is used to estimate a continuous surface of environmental measures (e.g. noise level) from point observations. The estimated measure at each location (target location) is calculated as a weighted average of the observations within a search window (or bandwidth). The weight of each observation is decided by applying a kernel function to the distance between the target location and that observation. The kernel function is typically a distance decay function with a maximum value when the distance is zero, and with a 0 value when the distance exceed the bandwidth. The formula of kernel smoothing is shown below, where \( K() \) is the kernel function, \( h \) is the bandwidth, \((X_i, Y_i)\) is the location of observation \( i \), and \( Z_i \) is the environmental measures of observation \( i \).

\[
\frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}, \frac{y-Y_i}{h}\right) z_i}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}, \frac{y-Y_i}{h}\right)}
\]

 Performing the kernel smoothing with massive observations from multiple users are extremely computationally intensive. Hence, a parallel kernel smoothing algorithm is implemented based on CUDA\(^9\) (Compute Unified Device Architecture) to exploit the computational power of GPUs. Multiple parallel threads are launched simultaneously, each of which estimates the measure at one location (one cell for the output raster). Each thread searches through each of the observations, calculates the weight of this observation to its cell, and outputs the weighted average of these observations as estimated measure of its cell. In this case, the 50-meter kernel bandwidth distance is also incorporated as the bandwidth of the kernel smoothing method, and the output is a GeoTIFF image, which is overlaid on top of Esri world street map for visualization purpose.

4 User Case Scenario

A noise mapping use case is investigated by collecting data of sound pressure and noise level using a mobile application. The application utilizes the microphone of a mobile device to

---

\(^9\)http://www.nvidia.com/object/cuda_home_new.html
measure sound pressure with noise level calculated in decibels (dB) based on the following equation (Hansen, n.d.; Maisonneuve et al., 2009):

\[
L_p = 10 \log_{10} \left( \frac{p_{\text{rms}}^2}{p_{\text{ref}}^2} \right) = 20 \log_{10} \left( \frac{p_{\text{rms}}}{p_{\text{ref}}} \right) \text{dB}
\]

where \( p_{\text{ref}} \) is the reference sound pressure level with a value of 0.0002 dynes/cm\(^2\) and \( p_{\text{rms}} \) is the measured sound pressure level. According to the World Health Organization Night Noise Guidelines (NNGL) for Europe\(^{10} \), the noise level is considered normal for night time level 55 db and day time level of 70 db. This calculated value is also be calibrated by users as the condition of both physical environment and the mobile device varies.

The mobile application also assigns a pair of geographic coordinates (in the format of latitude and longitude) to each measured value together with a timestamp. The update time interval for each recording is every 5 seconds. The recorded measurements are saved directly on the mobile device and we let users decide when to upload their data to the server, whether immediately during taking the measurements or afterwards. An example for the data format of the measurements is shown in Figure 3. Note that the measurements of other sensors on a mobile device can be included. Given the diversity of sensors on different devices, we use a flexible data management approach based on MongoDB.

\[\text{Figure 3: An example of recorded noise measurements saved on a mobile device}\]

In this user case scenario, we choose the campus of University of Illinois at Urbana-Champaign and its surroundings as the study area and ask the participants to go around the campus to collect the noise level measurements. We choose the campus of University of Illinois at Urbana-Champaign and its surroundings as the study area and ask the participants to go around the campus to collect noise level measurements. The user interface of the mobile application is shown in Figure 4, where users have options to record, upload and interact with noise maps. The mobile application is implemented as a background service on the device and, thus, when participants are walking in their vicinities, they are free to be engaged in other activities.

From a generated noise map, we can identify those spots at which the noise level exceeds such ranges. In Figure 5, we can examine the visualization of the noise exposure to an individual participant along her or his trajectory. At the current stage, we have not quantitatively

\(^{10}\)http://www.euro.who.int/_data/assets/pdf_file
estimated accumulated noise exposure, which will be taken into account in our future work. Figure 6 shows the noise map of a specified hour using a 50-meter kernel bandwidth, which is generated from the measurements uploaded by all of the participants during the 1-hour period. From the visualized results, we can identify those spots where the noise pollution occurs (in red color) within the specified hour. A new feature, to be evaluated, for providing in-depth information about what causes such noise pollution is to allow users to append descriptive text when they carry out monitoring using their mobile devices. Figure 7 is the noise map of the same hour but using 100-meter kernel bandwidth, which demonstrates the effects of choosing different sound decay distance.

Figure 4: The user interface of the mobile application
Figure 5: Noise mapping along the trajectory of an individual participant

Figure 6: The generated noise map using a 100-meter kernel bandwidth during a specified hour
Figure 7: The generated noise map using a 100-meter kernel bandwidth during a specific hour.
5 Conclusions and Future Work

The availability of a variety of affordable mobile sensors is fostering participation of volunteered citizens in sensing urban environments using mobile devices. By utilizing embedded GPS receivers to append geographic coordinates to sensor measurements, the collective efforts from participatory urban sensing activities can provide high-resolution spatiotemporal data for creating pollution maps of large cities. In relation to the big data collected from such crowdsourcing activities, CyberGIS provides high performance computing and participatory computing architecture to support scalable user participation and data-intensive processing, analysis and visualization (Ghosh et al., 2012).

In this paper, we present a framework that utilizes several components of the CyberGIS platform to facilitate volunteered citizens to engage environmental monitoring using mobile devices. This frameworks is intended to incorporate readings from the environmental sensors on the mobile device. As the availability of sensors varies on different devices, this framework chooses a MongoDB (without the requirement for a predefined schema) cluster for data storage. A MapReduce approach is used to filter and extract trajectories of each individual participant to visualize the pollution exposure. It is also used for dynamically generating pollution maps by aggregating the collected sensor measurements using a time window. The pollution maps are rapidly generated by using kernel method via paralleled GPU. In this study, we only demonstrate the functionality of the framework using the case for dealing with the geo-tagged and timestamped noise level measurements, which is collected from our dedicated prototype mobile application using the combination of integrated GPS receiver and microphone on mobile device.

This paper describes a novel framework that utilizes several key components of CyberGIS architecture to enable volunteered citizens to contribute to environmental monitoring using mobile devices. As the availability and types of sensors vary on different devices, this framework adopts a MongoDB (without the requirement for a predefined schema) cluster for data management. A MapReduce approach is used to filter and extract trajectories of each individual participant to visualize pollution exposure. It is also used for dynamically generating pollution maps by aggregating the collected sensor measurements based on specified temporal durations. The pollution maps are rapidly generated using a parallel kernel density estimation algorithm based on GPU. As a case study, we demonstrate the functionality of the framework for dealing with the geo-tagged and time-stamped noise level measurements collected from the mobile application that uses the combination of integrated GPS receiver and microphone on mobile device.

At the current stage, there are some limitations regarding the implementation of the framework. For example, the selection of the kernel method assumes the measured values stay the same within the kernel bandwidth, which may not be the case in real-world scenarios. Also, the kernel method may not be suitable for generating other pollution maps, for example, air pres-
sure. Therefore, some domain knowledge is required for future improvement of the framework. In relation to trajectory extraction for visualizing pollution exposure to individual participants, quantitative methods for estimating actual exposure need to be explored. Furthermore, we plan to acquire environmental measurements from pertinent government agencies to validate the results that are produced based on data from volunteered participants. Finally, the current MapReduce method relies on the MongoDB cluster, where Apache Hadoop is being explored to improve computational performance.

Acknowledgements

This material is based in part upon work supported by the U.S. National Science Foundation under grant numbers: 0846655, 1047916, and 1354329. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors are grateful for insightful comments on the earlier drafts received from Yan Liu, Anand Padmanabhan.

References


Abstract:

As students of cities, we know very little about how the current context of decreasing urban economies is understood by the people who live in such locales and are affected by
planning interventions. In this paper, we contribute to the larger literature around neighborhood change, specifically decline, by asking: what ways can we learn about residents’ sentiments and civic concerns in the shrinking city context?

To pursue this research question, we use the case of the post-industrial city of New Bedford, Massachusetts. We collected data on the attitudes and opinions of residents of New Bedford by conducting 1) a conventional content analysis of published planning meeting minutes 2) a sentiment analysis of geo-coded Twitter messages in the city.

Meeting minutes have long been a source of data for qualitative and planning researchers. We reviewed over 300 meeting minutes from the years 2007-2013 to understand the deliberations of civically engaged New Bedford residents. A 25% sample is analyzed using conventional content-analysis techniques, with two researchers independently coding and then comparing their codes in order to improve inter-rater reliability (Gaber and Gaber 2007).

We designed a database to collect data from the Twitter API. Our database stores, indexes and analyzes a continuous stream of data from the Twitter Decahose (a 10% sample of all Tweets) that are geo-tagged to specific locations. Most Twitter posts little to nothing to do with perceptions of place. We filter the results based on a text search for key terms that pertain to New Bedford, their neighborhoods, or other specific places therein. Then, the ongoing sample of tweets is directly fed into SPSS Modeler to enable a sentiment analysis of how residents perceive their communities, using the same approach employed in analyzing the meeting minutes.
Keywords: qualitative analysis, child and family policy, content analysis, Twitter, meeting minutes
Introduction

A large body of sociological and psychological research has established that concentrated poverty is detrimental children and families (Leventhal & Brooks-Gunn, 2001, 2011, Shaw & McKay, 1942). A separate body of planning and economic research has established that urban economies in the United States have followed different trajectories, especially in the post-recession years (e.g., Hollander, & Németh, 2011). Yet we know very little about how the current context of decreasing urban economies is understood by the people who live these locales. Moreover, while the internet and the Web have resulted in paradigm shifts across a number of domains, one in particular is that these medium have made it possible to find out about the experiences and expressions of a “vast pool of people” (Pang and Lee, 2008) Despite the fact that mediums like Twitter are used as a form of communication, most research has focused on the quantitative rather than qualitative possibilities of big data.

We contribute to the emergent literature around place and perception by focusing on qualitative data and asking: what ways can we learn about residents’ sentiments and civic concerns in the shrinking city context?

To pursue this research question, we use the case of the post-industrial city of New Bedford, Massachusetts. We collected data on the attitudes and opinions of residents of New Bedford by conducting 1) a conventional content analysis of published planning meeting minutes 2) a sentiment analysis of geo-coded Twitter messages in the city.

Literature Review
1. How are large social media datasets analyzed?

In approaching the question of studying the sentiment of large urban populations, we have to consider ways in which social media datasets have been analyzed. While there are some naysayers (Goodspeed 2013) who worry about the validity of studying these unconventional data sources, many others have latched onto the new medium with aplomb. Some researchers have employed social network analysis to explore the ways individuals interact with one another (Hansen, et al. 2009; Ediger et al. 2010; Catanese et al. 2010) or the ways that people follow links (Namata 2010), or combing content with user comments (Eugene Agichtein et al, 2008). Much of this research uses social media to understand group processes and properties (Tang and Lui 2010) but does little to fundamentally reveal what people think about places.

Important social science research has used massive social media datasets to advance social objective (Ediger, et al. 2010), the forecast shifts in the mood of users (Servi, et al. 2012), to enhance journalistic investigations (Diakopoulos et al. 2010), and to infer users locations from their Tweets (Mahmud et al. 2012). For our paper, we follow a tradition of those using social media to conduct opinion mining and sentiment analysis (Gokulakrishnan et al, 2012; Martineau, 2009; Meeyoung Cha et al, 2010).

2. What data sources and software tools have researchers used?

Researchers have used many social media datasets to explore a range of social science questions. Facebook was studied by Catanese et al, 2010 and many others have used Twitter (Servi et al, 2012; Bruns 2011; Gokulakrishnan et al. 2012; Mahmud et al. 2012; Dou et al. 2012;
Meeyoung Cha et al, 2010). Flickr (Cha 2009), Yahoo! Answers (Agichtein et al. 2008) and Foursquare (Mahmud et al, 2012) have also been subject to rigorous study.

For research exploring social network analysis, scientists have employed GraphCT, a Graph Characterization Toolkit (Ediger et al, 2010) and NodeXL (Hansen et al, 2009). Servi et al. 2012 used a program called Linguistic Inquiry and Word Count 2007 to study linguistic and word counts. Bruns (2011) employed Gawk for processing large data networks and Gephi for network visualization, and Diakopoulos et al. 2010 used Vox Civitas as a visual analytic tool.

For conducting sentiment analysis, IBM’s SPSS Modeler (and its predecessor Clementine) have been used by Pachamanova and Fabozzi (2014) and Dursun (2009), among others.

3. How have Urban Studies / Environmental Studies / Urban Planning used these approaches?

Empirical and theoretical work on the application of big data to urban planning and urban environments has been quantitatively focused. While theoretically, social media is composed of three key components: users, content, and metadata, much of the quantitative research exploits the user and metadata components. Content, which is clearly central aspect of social media, appears that to have been studied less from a planning perspective. Empirical quantitative research using Twitter in particular has focused on movement histories (Fujisaka, 2010), navigation assistance (Salim, 2012), crowd behavioral patterns (Shoko, Lee, and Sumiya, 2011), characterization of land use (Frias-Martinez, Vanessa, et al, 2013), health and place research (Ben-Harush, Carrola and Marsh, 2012), the relationship between user socio-economic status and geographic mobility patterns (Cheng, 2011) and epidemiological, crisis management
(MacEachren, 2011) and emergency preparedness (Merchant et al, 2010) modeling. There have been some qualitative inquiries into the application of social media to urban settings. These test social media user’s reactions to interventions, such as a planning project (Schroeter and Houghton, 2011) or child welfare campaign (Paek, Hove and Cole). Other research focuses on the potential for social media to engage citizens in urban governance (Fredericks and Foth). However there do not appear to be any empirical studies that investigate how people living in urban environments use social media to engage with urban issues.

Researchers recognize the centrality of sociability and verbal communication to social media. According to Ivanov (2013). “Social media are web sites as means of interactions among people in virtual communities (social networks) where people create, observe, share, exchange and comment content among them”. Some also have identified a need to incorporate empirical studies of the qualitative data inherent to social media communiques. Current research encourages content and sentiment-based analysis of posts (Pak and Paroubek, 2010) for its potential to provide a rich source of context for better understanding users (Cheng, 2011). However, a review of the research in this emerging area suggests that the application of qualitative analysis of social media user content and sentiment in urban research contexts is rare.

4. What is the state of child/family policy in shrinking cities?

In the 1990s, something dramatic occurred in America’s cities: concentrated poverty declined. But a study recently released by the Brookings Institution (2011) has shown a reversal of that trend, as the number of people living in extreme-poverty neighborhoods (i.e., those where greater than 40% of residents are poor) has grown by about one-third. Rises in concentrated poverty were driven in large part by economic malaise that accelerated with the fallout from the
2008 collapse of the American housing market and the ensuing foreclosures crisis. Regions like New England, the Upper Midwest, and the Sunbelt have been hit particularly hard with respect to vacant housing and neighborhood destabilization, but of note is that poor people in cities remain more than four times as likely as their suburban counterparts to reside in neighborhoods marked by concentrated poverty.

On many fronts, these trends are cause for alarm. In the case of children and families, a rather vast literature has documented links between neighborhood poverty and compromised well-being (but questions remain about the magnitude and nature of these links; see Leventhal, Dupéré, & Brooks-Gunn, 2009, for a review). Although much less is known about the consequences of increasing concentrated poverty for children and families (Leventhal & Brooks-Gunn, 2001, 2011), urban sociologists have long highlighted the social problems, such as adolescent delinquency, that arise when neighborhood disadvantage increases—specifically rises in poverty, single-parenthood, residential instability, and racial/ethnic heterogeneity (Shaw & McKay, 1942). Increasing disadvantage is thought to undermine neighborhood social organization by weakening community institutions, hindering their ability to monitor residents’ behavior and maintain order. Recent work in this tradition reveals that increasing poverty lowers a community’s ability to come together around shared values and norms regarding child rearing (or “collective efficacy”; Sampson & Morenoff, 2006), with potentially adverse implications for children’s development (e.g., Browning, Leventhal, & Brooks-Gunn, 2005; Sampson, 1997; Xue, Leventhal, Brooks-Gunn, & Earls, 2005). Although this work is informative, much less is known about how the current context of increasing neighborhood disadvantage—marked by growing poverty, vacancies, unemployment and depopulation—is associated with child and
family well-being or what types of policies are likely to mitigate the potentially unfavorable consequences for them.

However, urban scholars have done less to take urban context into consideration when theorizing about neighborhood effects. While scholars recognize that neighborhoods marked by concentrated poverty operate within larger urban settings, the dynamics between the neighborhood level processes and city-wide structural characteristics is relatively unexplored. Scholars usually treat the urban ghetto situated within a large economically diverse metropolitan area as the archetypical neighborhood of concentrated poverty. However, concentrated poverty occurs in a variety of urban settings, from, on one end of the spectrum growing major metropolitan areas such as New York and Los Angeles to, on the other end of the spectrum, shrinking former industrial cities of such as Detroit and Cleveland. It seems to make sense that policy approaches for children and families might differ depending on the city context. Thus, policy makers and researchers do not know much about neighborhood effects for poor people living in shrinking cities as compared to stable cities.

The U.S. Government has led many efforts to address poverty and economic decline around the National Stabilization Program, pumping millions of dollars into the most ravaged cities (U.S. General Accountability Office, 2010). While federal policy has been recognized as useful, the real business of managing neighborhood change has historically fallen to local governments and community development organizations (Frisch & Servon, 2006; Hoch, 2000; Innes & Booher, 2010;). What remains unclear is the extent to which child and family well-being are assessed and addressed in these efforts and what local residents attitudes to urban conditions are.
**Data and Methods**

We pursued the following method in an effort to explore similarities and differences in the sentiment and use of key words by citizens across vastly different contexts. We compare the frequency of key word use in the formal context of civic meeting minutes to the informal context of Twitter. The key words we focus on are those germane to civic life and in particular ones that are relevant to child and family life. The goal was to offer a cursory but systematic comparison of the pattern of the appearances of these key words and sentiments across contexts.

We reviewed 44 meeting minutes for four Committees in New Bedford for 2013 to understand the deliberations of civically engaged New Bedford residents. Meeting minute selection was driven by a convenience sampling frame: most city Committees did not have their minutes posted online. In an effort to examine a range of topics under consideration by city committees, we selected the City Council, Parks Committee, Zoning Board of Appeals and the Board of Health.

We also designed a database to collect data from the Twitter API. Our database stores, indexes and analyzes a continuous stream of data from the Twitter Decahose (a 10% sample of all Tweets) that are geo-tagged to specific locations. We designed custom software that pulls only those messages that are Tweeted from a geographic areas delimited as a rectangle around the city (see Figure 1). Using 70.89941601 (Longitude) and 41.759387677 (Latitude) for the Northeast Corner and -70.981963 (Longitude) and 41.591322004 (Latitude) for the Southwest corner, the rectangle encompasses the entire geographic extant of New Bedford, along with small slices of some neighboring communities. A total of 122,187 messages were collected during the period of February 9 – April 3, 2014.
Once we had acquired and organized these two datasets we proceeded to conduct a sentiment and keyword analysis on both.

Using IBM’s SPSS Modeler, we ran the Twitter messages through a sentiment analysis module. The sentiment analysis employs an internal sentiment dictionary, where we selected the “Opinions (English)” sub-dictionary. The SPSS Modeler Text Analytics function provides many complex and nuanced ways to study language. For this research, our aims were modest and we selected a relatively simple dichotomy to guide our analysis: all messages were coded as either positive, negative, or N/A by the Modeler software based on the appearance of certain sentimental words. Another commonly used sentiment dictionary is known as AFINN, developed by Finn Årup Nielsen has 2,477 words and phrases, with each one rated on an ordinal scale of +5 to -5 (e.g. obliterate is -2, where rejoicing is +4) (Hansen, et al. 2011; Ngoc and Yoo 2014). Based on discussions with IBM officials we feel comfortable that the Modeler dictionary likely bears some resemblance to the AFINN dictionary – though the exact parameters of the Modeler product are proprietary. For additional validation, we also ran the Twitter messages through the AFINN dictionary using custom-designed software.

We next ran the same sentiment analysis on the 2013 (n=44) meeting minutes and then did an old fashioned human-based content analysis on just one set of meeting minutes. This additional triangulation allowed us to provide that additional check on context and how the sentiment analysis was functioning.

Next, we sought to better explore the way that the discourse in each medium engaged with the topics of urban policy. While City Committee meetings would be expected to cover such topics, Twitter users’ posts would not be expected to relate to perceptions of place. We
conducted a keyword text search for a set of 24 terms that pertain to New Bedford, their neighborhoods, or other topics related to child and family policy in a shrinking city.

Through a grounded approach we developed a set of 24 keywords for the qualitative analysis. The keywords were developed through a close reading of the meeting minutes. A research assistant reviewed the meeting minutes of the School Committee, Zoning Board, Planning Commission, City Council, Board of Health and City Council from the years 2008-2014 and wrote a summary narrative for each year. The researchers reviewed the narratives and the research team identified 24 frequently appearing terms relevant family and child development.

The researchers used Nvivo qualitative software to analyze the 2013 meeting minutes. Researchers chose this time frame because it represented the most current time frame that had the most complete notes. The meeting minutes (44) analyzed included the City Council (20), Park Board (11), Planning Board (2), Zoning Board (7) and Board of Health (5). Researchers classified the 24 keywords as nodes. The software allows for searching by nodes, providing an account of the context the key word appears in, as well as an accounting of the number of times the keywords appear.

The keyword search was done by hand (Control-F in Microsoft Word) in a search through all of the Twitter posts.

For both analyses, we ran basic descriptive statistics and frequencies, and then ran means tests to examine how similar the two datasets were.
Results

For six weeks from February 2014 to April 2014, we used a custom-designed software “Urban Attitudes” to download 122,186 messages. Of those, 87,079 had valid fields and were tested for sentiment. The messages were collected in a relatively uninterrupted 24/7 basis into a series of individual *.csv files. We compiled those *.csv files into a single file and ran it through the SPSS Modeler Text Analytics program. The result was that 6,268 (7.2%) of the messages were classified as Positive (including several variations of the concept of Positive, including Positive Attitude, Positive Budget, Positive Competence, Positive Feeling, Positive Feeling (Emoticon, ☺), and Positive Functioning. A total of 4,825 (5.5%) were classified as Negative (along with conceptually close variations of negative).

Anyone familiar with Twitter might not be surprised at how few messages evoked any sentiment, a total of 87.3%. Mostly, people use Twitter to communicate about their favorite sports team, what they are having for dinner, or their plans for the evening. But, these results show that over 10,000 messages did have an embedded sentiment, and it was significantly more positive than negative.

To check these results, we also ran the Twitter messages through a custom-designed sentiment analyzer that used AFINN. Here the unit of analysis is the word, where the Modeler software used the Tweet as the unit of analysis. Using the AFINN dictionary and none of contextual information embedded in the Modeler software, we found 58,490 positive words and 44,981 negative words (among a total word count of 1,139,761). Presented as a percentage, that means 5.1% positive words and 3.9% negative words, not too different than the Modeler results. Each AFINN word is weighted, so with the added weights (between -5 and +5), the positive...
words had a combined score of 132,838 and the negative words had a weighted combined score of -111,529.

For the 44 meeting minutes, Modeler found the appearance of 326 positive words and 130 negative words (using the same set of allied concepts presented earlier). The meeting minutes comprised of a total of 19,433 words, thus the positive words accounted for 1.7% and the negative words were only 0.7%.

Turning now to the keyword search, we did a by-hand search through all of the meeting minutes to determine the number of appearances of our list of 24 keywords (see Table 1). Here we found that the words “schools”, “zoning”, “health”, “safety” and “parks” appeared most frequently (see Table 2). Some words which had been important in civic deliberations, like truancy, did not appear at all in the 2013 meeting minutes notes. A close reading of the notes indicated that the city was engaged in a long-term planning project regarding the rebuilding of a school, implementing a program to improve children’s health and fitness and attempting to respond to national level issues such as gun control and medical marijuana legalization. Likewise, we searched through the complete Twitter message compilation, counting the appearance of those same keywords. See Table 2 for the results.

Conclusion

In this paper we took an initial step toward contributing to emergent literature around place and perception by asking the following question: *what ways can we learn about residents’*
sentiments and civic concerns in the shrinking city context? We found several possible answers to these questions. The first is that in the case of New Bedford residents rarely expressed either positive or negative sentiments in either their formal civic deliberations and informal exchanges on Twitter. A close reading of the meeting minutes suggested that the lack of sentiment can be partially explained by the fact that the meetings were largely procedural events, rather than a spirited exchange of ideas and opinions one associates with deliberative democracy. However, with respect to Twitter, given that it is a medium that seems to nearly compel users to be emotionally expressive (i.e. to use emoticons) it is somewhat surprising that the measured levels of expressiveness suggest a certain stoicism on the users part.

The second finding is that there is some overlap between the rankings of how frequently Twitter posters and meeting attendees use key words that reflect urban concerns with child and family issues. The top ranking keywords from the meeting minutes were often also the top ranking key words in the Twitter. The words “school,” “health,” “safety,” “parks,” “field,” and “children” were high ranking key words in both analyses. While one should not lose sight of the fact that these words very rarely appeared in either data set, the similarity across rankings is somewhat interesting. It suggests the possibility that across formal and informal contexts residents express similar levels of interest in civic issues. Discussions of health, safety, parks and children (field and school have too many alternative uses i.e. “field” also refers to a sports field) arise with somewhat parallel frequency.

Although this analysis is very preliminary, it is also intriguing. We collected the meeting minutes for a short period of time -- one year -- and the Twitter feed for an even shorter period -- four weeks. This does not allow for a truly robust analysis, especially given the ambitions of “Big Data” analysis. It is possible that any dataset would contain the key words at similar
frequency and this is simply a reflection of English language usage. However, it could be that the pattern identified above reveals in both formal and informal settings residents engage in topics in similar ways. That would suggest that the informal world of communication is an extension of, not necessarily an alternative to, formal communication and vice versa. However, given the limitations of this analysis, this is entirely speculative. Reviewing the data across larger time frames and comparing it to other urban settings might help test these ideas further.

Additionally, the obvious question is, how would one classify the majority of Twitter communications, since it appears that in the case of New Bedford in 2014, most could not be classified as either positive or negative. A conventional content analysis might help illuminate how people use Twitter to communicate and how sentiment fits into this modern means of expression.

References


Gaber, John and Sharon Gaber. 2007. Qualitative analysis for planning & policy: Beyond the numbers. Chicago: APA Planners Press.


Figures and Tables
Figure 1: City of New Bedford, enveloped by a rectangle to facilitate downloading of Twitter messages.
1. Children
2. School
3. preschool
4. Safety
5. Vacant lot/vacancy
6. Housing violation
7. Field
8. Truant
9. Redevelop
10. Parks
11. Guns
12. Alcohol
13. Underage drinking
14. Violence
15. Foreclosure
16. Zoning
17. Health
18. Demolish
19. Smoke/smoking
20. Tobacco
21. Condemned
22. Lead
23. Prenatal
24. Health

Table 1: Keywords used in meeting minute and Twitter messages searches.

<table>
<thead>
<tr>
<th>Keywords</th>
<th>minutes</th>
<th>twitter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#</td>
<td>rank</td>
</tr>
<tr>
<td>School</td>
<td>346</td>
<td>1</td>
</tr>
<tr>
<td>Zoning</td>
<td>227</td>
<td>2</td>
</tr>
<tr>
<td>Health</td>
<td>140</td>
<td>3</td>
</tr>
<tr>
<td>Safety</td>
<td>121</td>
<td>4</td>
</tr>
<tr>
<td>Parks</td>
<td>75</td>
<td>5</td>
</tr>
<tr>
<td>Housing violation</td>
<td>67</td>
<td>6</td>
</tr>
<tr>
<td>Smoke</td>
<td>42</td>
<td>7</td>
</tr>
<tr>
<td>Field</td>
<td>39</td>
<td>8</td>
</tr>
<tr>
<td>Children</td>
<td>35</td>
<td>9</td>
</tr>
<tr>
<td>Vacant</td>
<td>25</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 2: Appearance of Keywords in Meeting Minutes and Twitter messages (Top 10)
The designer as regulator – design patterns and categorization in citizen feedback systems

Dietmar Offenhuber, Assistant Professor
Dept. of Art and Design / School of Public Policy and Urban Affairs
Northeastern University
d.offenhuber@neu.edu
Abstract

The relationship between emerging disciplines focusing on ‘Big Data’ and Urban Planning research has not always been an easy one. In the recent decade, new data-intensive methods of inquiry have taken shape and are rapidly gaining influence on the studies of cities and social systems, yet (sometimes fundamental) differences and tensions remain around issues such as data-driven versus theory-driven forms of inquiry, the merit of prediction versus the need for causation, and finally questions of politics and hidden biases of data. While the current debate between both enthusiasts and critics of ‘Big Data’ is dominated by concerns about “truth” and “bias”, I argue for a more differentiated view. Using the example of the categorization schemes of citizen-reporting systems, I argue that we need to take the design of systems and interfaces as a behavior-shaping and constraining factor into account. This paper identifies design-patterns that regulate the interaction between users of citizen feedback apps and cities and investigates the ways in which categorizations used in the interface correspond to the reports submitted by citizens.

Keywords: Citizen Feedback Systems, Categorization, Design Implications

Introduction

The issue of bias is at the core of the discussion around the use of big data in social and urban studies, both stressed by enthusiasts and critics of data-intensive approaches to research. Computational Social Science makes heavy use of telecommunication data to study human mobility or online search behavior and social media for demographic research (Lazer et al., 2009; Saiz & Simonsohn, 2007). Although ultimately generated by human behavior, these opportunistic data sources were generally not collected for research purposes and are subject to all kinds of hidden variables and biases. Data-intensive methods are built on the assumption that large quantities of data allow controlling for the multitude of unobservable variables, thus outweighing the disadvantages created by the non-probabilistic nature of these data sources (Saiz & Simonsohn, 2013). Inspired by the explosive growth of large data sources that are availability at almost no cost, Chris Anderson proclaimed “the end of theory” (Anderson, 2008), echoing the notion by the late Jim Gray that data-driven methods will supersede the primate of empirical,
analytical, and simulation approaches and become the basis for a ‘fourth paradigm of
data-intensive scientific inquiry’ (Hey & Tolle, 2009).

In the light of this enthusiasm, Computational Social Scientist David Lazer
diagnosed “Big data hubris”, evident in the widespread idea that predictions based on the
analysis of social media can substitute more traditional approaches to research (Lazer,
Kennedy, King, & Vespignani, 2014). Critics of “Big Data” emphasize the nature of data
as cultural and inherently political artifacts rather than a neutral resource (Gitelman,
2013) or doubt that the sheer volume of data can make the often unknown biases
inscribed in the data disappear (Boyd & Crawford, 2012).

Lazer’s analysis of ‘Google Flu Trends’ revealed that the service analyzing billions of
Google search queries for predicting flu outbreaks performs worse than traditional
models based on ‘small data’. Furthermore, its results were sensitive to a number of
external factors such as the changing designs of the search algorithm and the resulting
behavior changes of the users trying to hack the system (Lazer et al., 2014).

On the surface, this seems like a straightforward issue of bias – after all, we can
compare the prediction to an observable outcome. However, if we want to understand the
dynamics between search algorithm and user behavior, a truth / bias dichotomy is
problematic: there is no stable baseline in terms of user behavior and we have to deal
with the constructed nature of data itself. The same is true for all data collected from
users of communications technologies. Communication is highly contextual, the
properties of the medium shape the way how people use it (Short, Williams, & Christie,
1976). A purely data-centric view is always after-the-fact, and ignores the context of the
data collection process as well as the role of the interfaces and system designs that
influence collected data in non-obvious ways. Paraphrasing Jay Forrester, we cannot
choose not to design, we only have the choice between different designs (Forrester,
1970).

Citizen feedback systems as a case study

This paper aims at shifting the focus of the big data discussion away from the truth/bias
dichotomy, towards a discussion of design and its implications for urban governance.
Based on the example of citizen feedback systems, this paper investigates how the design
of an interface influences the data that are collected through it. By identifying design patterns for citizen feedback systems, this paper shows how design guides user interaction and shapes the conversations and reports submitted through the system.

During the past 15 years, most cities have established a unified infrastructure for receiving citizen requests and complaints, often called 311 systems after the FCC designated 3-digit dial code for non-emergency services. The original telephone hotlines were supplemented through other communication channels; today almost every major US city offers a smart phone application for submitting reports. Currently, over 100 applications exist: clients and web-services exist for submitting citizen reports, designed by cities, private companies, NGOs or by amateurs during hackathons or fellowships.¹ The evolution of these tools is based on ad-hoc prototypes and experiments. Nevertheless, most systems offer a similar functionality: the user can submit a textual report of an incident, attach a GPS location and optionally a photo of the incident (for example a pothole or broken streetlight), or view a list of reports submitted by others in the area. While the value of such feedback is generally accepted today, this was not always the case. Data collected by volunteers and citizens were met with concerns regarding their accuracy and actionability. However, a number of studies have shown that volunteer reported incidents are often more accurate than authoritative data, which is not always accurate either (Van Ryzin, Immerwahr, & Altman, 2008). Cities that have established 311-type systems connect a number of expectations with these systems: they should improve citizen participation, as well as the responsiveness of the city to citizen concerns. They should simplify the interaction between citizen and city, and at the same time generate actionable data for urban maintenance. Finally, these systems should help establishing accountability – even if this means being publicly shamed through submitted reports, which are usually publicly visible.

This paper is based on the hypothesis that all of these – sometimes aligned, in other cases incompatible – expectations are reflected in the design of these systems, intentionally or not. The effectiveness of advocacy by shaming through incident reports, for example, depends on the extent to which degree makes these reports not only public, but also visible and discoverable. Beyond data collection, citizen feedback apps change

¹ See a constantly updated list at http://wiki.open311.org/GeoReport_v2/Support
how citizens interact with cities and thus play an increasingly important role in urban governance. Modeled after message boards or twitter feeds, the systems facilitate a sustained conversation between reporter and the city instead of isolated interactions, and make this conversation publicly visible. Physical spaces are increasingly regulated through digital interfaces, as Martin Dodge and Rob Kitchin demonstrated with the examples of airports, domestic spaces and spaces of consumption (Dodge & Kitchin, 2004). As human behavior in urban space is shaped and constrained by these interfaces, system and interface designers assume, often unknowingly, the role of policy makers and regulators – their systems shape human behavior, constrain or facilitate interaction, and manage visibility of data artifacts.

This paper builds upon the initial results of a recent study, in which I investigated the differences between two smartphone citizen report apps used in the city of Boston – namely the city’s own system, CitizensConnect (CCN)² and the private system SeeClickFix (SCF)³ (Offenhuber, 2014). Since both systems are not only integrated with the city’s constituent relationship management system (CRM) operated by the same city department, but also use the same open311 protocol and therefore produce data in the same format, they an ideal case for comparison of the interface effects with everything else being equal. While both systems offer similar functionality and offer the same data collection protocol Open311 (OpenPlans, 2012), they show important subtle design differences which I linked to the different understandings of the system developers and providers. The city-initiated system CCN emphasizes service delivery, responsiveness and actionability of reports, while SCF emphasizes social accountability and a broader discussion around civic issues.

Using a grounded-theory content analysis of the textual descriptions of a randomly drawn sample of 2200 reports submitted to each respective system, the study revealed consistencies, but also important differences. First, the two systems were used to discuss different types of issues. Graffiti reports played a prominent role in CCN, but were almost completely absent in SCF. Discussions on SCF on the other hand, frequently revolved around traffic issues, infrastructure repair and governance issues. But also the

³ [http://seeclickfix.com/boston](http://seeclickfix.com/boston)
style and rhetoric was different between the two systems: while CCN reports framed issues frequently as personal nuisance, SCF reports often conveyed the urgency of the incident by stressing its implications for the larger public. SCF reports generally exhibited a more critical tone compared to CCN. CCN on the other hand, included more reports concerned the behavior of other people. Figure 1 lists the main differences between the identified qualities and features in both data sets. It can be speculated about what causes those differences – the way its developers advertise the system, the culture of the user-base, but also the design of the interface in relation to the goals and prior assumptions of its developers.

<table>
<thead>
<tr>
<th>Service category</th>
<th>N</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>573</td>
<td>49%</td>
</tr>
<tr>
<td>Graffiti</td>
<td>201</td>
<td>17%</td>
</tr>
<tr>
<td>Pothole</td>
<td>175</td>
<td>15%</td>
</tr>
<tr>
<td>Streetlight</td>
<td>122</td>
<td>10%</td>
</tr>
<tr>
<td>Sidewalk Patch</td>
<td>41</td>
<td>3%</td>
</tr>
<tr>
<td>Damaged Sign</td>
<td>25</td>
<td>2%</td>
</tr>
<tr>
<td>Unshoveled Sidewalk</td>
<td>26</td>
<td>2%</td>
</tr>
<tr>
<td>Roadway</td>
<td>9</td>
<td>1%</td>
</tr>
<tr>
<td>Total</td>
<td>1172</td>
<td>100%</td>
</tr>
</tbody>
</table>

*Figure 1 Citizens Connect: left: number of service requests by service category; right: an example of a report and the reply from the city (Offenhuber, 2014).*

<table>
<thead>
<tr>
<th>Incident type</th>
<th>CCN N %</th>
<th>SCF N %</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animals</td>
<td>16 1%</td>
<td>7 1%</td>
<td></td>
</tr>
<tr>
<td>Graffiti</td>
<td>210 18%</td>
<td>16 2%</td>
<td></td>
</tr>
<tr>
<td>Ice</td>
<td>37 3%</td>
<td>3 0%</td>
<td></td>
</tr>
<tr>
<td>Infrastructure improvement</td>
<td>39 3%</td>
<td>92 13%</td>
<td></td>
</tr>
<tr>
<td>Infrastructure repair</td>
<td>493 42%</td>
<td>466 67%</td>
<td></td>
</tr>
<tr>
<td>Other violation</td>
<td>40 3%</td>
<td>13 2%</td>
<td></td>
</tr>
<tr>
<td>Plants</td>
<td>42 4%</td>
<td>11 2%</td>
<td></td>
</tr>
<tr>
<td>Social issues</td>
<td>7 1%</td>
<td>10 1%</td>
<td></td>
</tr>
<tr>
<td>Test / unknown</td>
<td>13 1%</td>
<td>8 1%</td>
<td></td>
</tr>
<tr>
<td>Traffic</td>
<td>64 5%</td>
<td>39 6%</td>
<td></td>
</tr>
<tr>
<td>Trash / litter</td>
<td>211 18%</td>
<td>29 4%</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1172 100%</td>
<td>694 100%</td>
<td></td>
</tr>
</tbody>
</table>

*Figure 2 Comparison Citizens Connect (CCN) vs. SeeClickFix (SCF): Types of incidents submitted (left), Motivations expressed (right) after a manual content analysis of 2200 reports (Offenhuber, 2014).*
Figure 3 Comparison Citizens Connect (CCN) vs. SeeClickFix (SCF): Tone of report (left), selected non-exclusive Properties (right) after a manual content analysis of 2200 reports (Offenhuber, 2014)

<table>
<thead>
<tr>
<th>Tone of report</th>
<th>CCN N</th>
<th>CCN %</th>
<th>SCF N</th>
<th>SCF %</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Critical</td>
<td>205</td>
<td>17%</td>
<td>204</td>
<td>29%</td>
<td></td>
</tr>
<tr>
<td>Friendly</td>
<td>54</td>
<td>5%</td>
<td>30</td>
<td>4%</td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>577</td>
<td>49%</td>
<td>317</td>
<td>46%</td>
<td></td>
</tr>
<tr>
<td>No text</td>
<td>234</td>
<td>20%</td>
<td>104</td>
<td>15%</td>
<td></td>
</tr>
<tr>
<td>Plea</td>
<td>64</td>
<td>5%</td>
<td>20</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td>Very critical</td>
<td>38</td>
<td>3%</td>
<td>19</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1172</td>
<td>100%</td>
<td>694</td>
<td>100%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Properties (non exclusive)</th>
<th>CCN N</th>
<th>CCN %</th>
<th>SCF N</th>
<th>SCF %</th>
<th>difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demanding Accountability</td>
<td>60</td>
<td>5%</td>
<td>30</td>
<td>4%</td>
<td></td>
</tr>
<tr>
<td>Complaint in strong language</td>
<td>37</td>
<td>3%</td>
<td>21</td>
<td>3%</td>
<td></td>
</tr>
<tr>
<td>Reporting other people</td>
<td>118</td>
<td>10%</td>
<td>26</td>
<td>4%</td>
<td></td>
</tr>
<tr>
<td>Concern for safety</td>
<td>143</td>
<td>12%</td>
<td>140</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Suggesting Improvements</td>
<td>53</td>
<td>5%</td>
<td>81</td>
<td>12%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4 Comparing normalized word frequencies of common words in reports submitted to CCN and SCF. The size of the dots gives a sense of the overall frequency of the term in both systems combined. It can be noticed that terms related to graffiti, trash, parking are more frequently occurring within CCN, while potholes and traffic issues are proportionally more frequent in SCF.

**Scope & research methods**

The overall research question of this paper asks: in what ways do the arrangements of the interface influence reporting behavior? And more specifically: how do submitted reports correspond to the categorization of incidents offered by the interface?
Based on the insights from the previous study, this paper will first discuss specific design patterns relevant for citizen feedback applications, and then focus on specific design pattern related to categorization schemes and knowledge structures. Expanding from the previous study, this paper will consider existing citizen feedback applications in the whole US and use machine learning techniques to investigate a larger data volume of over 65,000 reports collected via CCN and 5000 reports collected via SCF in the larger Boston area.

Manual content analysis is well suited for a differentiated qualitative investigation of textual data, but has limitations in terms of scale. Only a small subset of the data can be investigated due to the time-consuming iterative coding process. Automatic text analysis is not bound to scale issues, but comes with its own limitations. Due to the short length and contextual nature of reports, pattern extraction and their interpretation is challenging. Consequently, tone and intent of incident reports are not subject of this study.

Using the results of manual analysis as a baseline, I use automated text analysis on the whole corpus of reports. Probabilistic topic modeling is a computational method for discovering and annotating themes in large collections of documents of arbitrary size. A ‘topic’ is defined in this case as a distribution of fixed words over a document (Blei, 2012); a salient group of words that tends to co-occur across a single document. The most frequently method for topic extraction is the Latent Dirichlet Allocation (LDA) Algorithm allows the discovery of topics without prior annotation of documents (Blei, Ng, & Jordan, 2001).

**Design patterns**

Design is a broad and slippery term. While most people would associate Interface Design with the visual arrangements of interactive elements on the screen, the choice of color pallets, and the overall structure of the user experience, it includes much more. Since knowledge structures such as service categories are a central component of the user experience, categorization should be considered part of the design process. Design patterns are a common way to operationalize the broad and fuzzy-edged concept. Design patterns are a heuristic, structured approach to development involving a set of
abstract reusable templates, influenced by the architect’s Christopher Alexander’s ‘Pattern Language’ (Alexander, Ishikawa, & Silverstein, 1977; Gamma, Helm, Johnson, & Vlissides, 1995). Most computer interfaces are instances of design patterns, which offer strategies for data management, code development and user interaction. Patterns are domain specific, therefore it cannot be taken for granted that familiar social software patterns including user profiles, share and like buttons, comment sections, many-to-many communication and so on, will work equally well in the domain of urban apps and citizen feedback systems (although many of the investigated systems take precisely this approach).

The ad-hoc development history of civic technologies has not yet led to established software design pattern for urban applications, nevertheless, a few themes can be observed:

1. Affordances of the interface: which channels are available for users to submit their report?
2. Visibility: how is the visibility of submitted reports managed?
3. Social presence: how are users made aware of the presence of other users on the system?
4. Categorization: How is the classification of user complaints managed in the interface?

I will briefly discuss each pattern before entering a more in-depth discussion of categorization.

Affordances of the interface

The interface frames communication through the offered of expression. While existing interfaces do not utilize the expressiveness of spoken language to clarify the issue and its urgency, they have other advantages. Websites and smartphones can determine the location of the reporter in varying degrees of accuracy, which eliminates the need for verbal explanations. Users of smartphone applications have additional sensors at their disposal. Besides a better location sensing through GPS, smartphones also allow taking a
picture of the incident and attaching it to the report, often make no further explanation necessary. As a result, smartphone reports are highly contextual, which makes textual analysis of the report challenging. The interface also allows following up on previously submitted reports, creating a tendency towards sustained interaction instead a single comprehensive reports. However, textual input on smartphones is severely limited. Longer texts are tedious to write with on-screen keyboards, making reports often ambiguous.

CCN is centered on a mobile smartphone application (the city has other mechanisms using different modalities). While it is possible to submit a CCN report over the website, the system is rudimentary and not featured on the city’s website. SCF offers both web and mobile interfaces, but the website has more prominent role, resulting in longer reports. The average report on SCF has a length of 187 characters (median 125) compared to 87 for CCN (median 58). These affordances likely have consequences for the reporting of incidents. A smartphone user might be more willing to report an issue that requires less textual explanation.

Visibility of reports

Once a report is submitted, the question arises who can see the report (besides the official responding to the report). Almost all citizen feedback systems publish their reports publicly, but with different degrees of visibility – sometimes available in real-time and linked on the homepage of the city, sometimes only available upon request. Management of visibility is a complex issue. High visibility, i.e. by featuring single reports on the city’s homepage, can encourage participation and motivate citizen to use the system, which is desirable for the city. A high public visibility affords, however, also opportunity for vandalism by submitting offensive or unrelated content. High public visibility is also an incentive to hijack the system for pursuing special interests, for example by political opponents. While participation is desirable for the city, these described negative effects are likely not. Since manually moderating the visibility of reports is for most cities out of the question, the city has to manage visibility though design of the interface. Again, we can observe an important difference between mobile and full websites. The small screen

4 [https://mayors24.cityofboston.gov:4443/reports/list_services](https://mayors24.cityofboston.gov:4443/reports/list_services)
of smartphones limits visibility and makes it difficult to search for reports. The larger real estate of a laptop or desktop computer screen allows more sophisticated interfaces for browsing, searching and comparing reports. But the smartphone offers other advantages: by using location of the user as a filter, data can be shown in a more contextual way.

CCN’s web interface, following the template of twitter, shows a constantly updated stream of submitted issues. It is transparent, yet opaque, since it is difficult to find past reports and compare, filter and aggregate past reports involving a particular issue and subject. Being more web-centric, SCF keeps issues in a thread and are more persistent, resembling an Internet forum. Users can filter reports by location and can create watch areas – areas with a user-defined perimeter that become a filter for the stream of reports. In addition, the contents of SCF reports are indexed by search engines.

Social presence

The term “social presence” describes to which degree a communication medium is able to convey the verbal and non-verbal expressions of a partner in a conversation (Short et al., 1976). Social Presence is an affordance of a medium. The telephone can convey qualities of speech such as tone, but fails at conveying gestures. Short, Williams and Christie hypothesized that users of a particular telecommunication medium are aware the social presence this medium affords, and consequently adapt their communication in response. We communicate differently over SMS, email, or telephone. This also means that social presence can be influenced by design, and design can be used to regulate the way in which people communicate using this system.

The compared smartphone applications SCF and CCN offer a similar set of functionalities. Their most salient differences concern social presence. Most importantly, SeeClickFix allows other users to comment on submitted issues, which can create a strong sense that the reported issue is a concern for a large number of people. SCF encourages users to create self-descriptive profiles, and shows them in user interfaces as neighbors.

5 City of Boston provides a different online tool for this type of analysis, which however does not offer access on the textual content level
Classification schemes and reporting behavior

Classification schemes can tell us a lot about organizations, their politics and practices. For citizen feedback applications, incident categories are a central component that negotiate between the ways in which users perceive an incident and the formats that the city’s organization can process and act upon. In this capacity, categories frame the discussion between citizens and the city: they encourage certain types of reports, and discourage others. As Susan Star and Geoffrey Bowker observed, “Each standard and each category valorizes some point of view and silences another” (Bowker & Star, 2000, p. 5). Categories allow nudging users towards submitting actionable reports, which is in the interest of city and provider. A complaint about a damaged trash bin is more actionable than a complaint that the park should look prettier. The latter example might be better suited for starting a broader conversation, but the first is an issue that can be added to the work queue of the department of sanitation.

Currently, there is no default taxonomy for 311 systems, and each city has a different understanding of what set of incident categories are needed depending on the specifics of the city, the organizational structure of city departments, or historical contingencies. This does not necessarily have to be that way, comparable initiatives such as 211 dial code for getting information about human services have been formalized in a comprehensive taxonomy (LA County, 2014). The open311 protocol does not offer default incident categories, although the issue has been discussed on the developer list.6

Most cities do have a centralized catalog of service and request categories. However, the way city departments categorize and process incidents may seem overly complicated and opaque to the citizens, who do not know and care about organizational boundaries and cannot be expected to articulate their concerns in a way the city can use immediately. For 311 telephone services, the call center operators categorize the caller’s concerns into an appropriate service code used by the city. For requests submitted via website or mobile apps, this option does not exist, and system designers have to find the best mapping between user perceptions and internal categorizations.

6 See for example http://lists.open311.org/groups/discuss/messages/topic/1r3JjU001kyvQhceBpKbTK/
Different cities have chosen different solutions to deal with this problem. A quick survey of places using citizen feedback systems shows a multi-faceted picture. Using the open311 server directory\textsuperscript{7,8} and SeeClickFix place definitions queried through the Open311 API list service request,\textsuperscript{9} I have compiled a data set of 157 cities and their individual service categories.

The largest number of categories is offered by New York City’s system, accepting 1776 distinct service categories. A user submitting a request via the website can traverse a hierarchy until the appropriate service category is found. Since this option is impractical on the small screen of a smart phone, the most recent version of the NYC311 app (June 2014) offers a small subset of 20 categories.\textsuperscript{10} Also Bloomington and Washington DC use a large set of categories; 50 and 84 service types, respectively. However, most cities have settled on a much smaller number, with a median of five categories (12 of 118 cities with more than one categories). Especially smaller cities only offer a general category; this is the case for 38 of the 157 systems.

\textsuperscript{7} \url{http://wiki.open311.org/GeoReport_v2/Servers}
\textsuperscript{8} \url{http://open311status.herokuapp.com/}
\textsuperscript{9} \url{http://dev.seeclickfix.com/v2/places/}
\textsuperscript{10} \url{http://www1.nyc.gov/connect/applications.page}
Figure 5 Distribution of the number of service categories in 157 citizen feedback systems, color specify the number of systems with / without a general (‘other’) category.

Besides the number of categories, the specificity of these service descriptions varies across the data set. General, catchall categories (‘other’) are common throughout the data set, but also departmental assignments (‘Albany housing authority issues’), location-dependent issues (‘issue at fresh pond’), and specific issue categories (‘parking meter problem’) can be found. After gentle data cleaning (combining service names such as ‘potholes’, ‘pothole’, ‘pothole repair’ into one category),

Table 1 below lists the most frequent categories from the list of the 118 cities that have more than one category. Garbage related issues are underrepresented, as they appear as different service categories such as ‘litter’, ‘garbage’, ‘waste’, ‘trash,’ in varying contexts that are too different from each other to justify combining them into a single category.
Table 1 most frequently used service categories

<table>
<thead>
<tr>
<th>Rank</th>
<th>Incident / Service Category Type</th>
<th># Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pothole</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>Other</td>
<td>80</td>
</tr>
<tr>
<td>3</td>
<td>Street light</td>
<td>76</td>
</tr>
<tr>
<td>4</td>
<td>Graffiti</td>
<td>74</td>
</tr>
<tr>
<td>5</td>
<td>Sidewalk</td>
<td>39</td>
</tr>
<tr>
<td>6</td>
<td>Tree</td>
<td>29</td>
</tr>
<tr>
<td>7</td>
<td>Abandoned vehicle</td>
<td>27</td>
</tr>
<tr>
<td>8</td>
<td>Illegal dumping</td>
<td>26</td>
</tr>
<tr>
<td>9</td>
<td>Traffic signal</td>
<td>24</td>
</tr>
<tr>
<td>10</td>
<td>Dead animal</td>
<td>19</td>
</tr>
<tr>
<td>11</td>
<td>Street sign</td>
<td>16</td>
</tr>
<tr>
<td>12</td>
<td>Street sweeping</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>Animal control, Bucket request, Business requests, Comments and complaints, residential requests</td>
<td>8</td>
</tr>
<tr>
<td>14</td>
<td>Parking enforcement, Parks</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>Litter, Property maintenance, Street repair, Trash</td>
<td>5</td>
</tr>
</tbody>
</table>

Even after removing cities with only a single general category, ‘other’ remains one of the most frequently offered service types. If the ‘other’ category is not offered, users are forced to choose a certain type of report. An extreme example is the city of Toronto, which offers six service types. Five of these six categories concern graffiti: ‘Graffiti on a City road’, ‘Graffiti on a City bridge’, ‘Report minor pothole damage’, ‘Graffiti on private property’, ‘Graffiti on a City sidewalk’, ‘Graffiti on a City litter bin’.

Also Chicago and San Francisco offer a limited number (13 and 14, respectively) of specific service categories without offering a general category. However, users seem to appreciate the availability of a general category: in Boston’s CCN system, ‘other’ reports account more than half of all submitted reports (53%). Comparing service types across observed systems reveals other peculiarities. The city of Darwin divides service categories into the two broader categories ‘Reports’ and ‘Requests’. Baltimore offers a category for ‘City Employee Praise’ as well as one for ‘City Employee Complaint’.

Surprisingly, only the city of Bloomington provides a category for suggestions, despite numerous incident reports contain suggestions rather than incidents (Offenhuber, 2014).

How do service categories influence reporting behavior?

In September 2011, SCF announced a partnership with the City of Boston, which resulted in an integration of SCF with the city’s own CCN for reports submitted within the Boston
area. Until then, SCF did not prescribe categories; users could choose a custom headline for summarizing the reported incident. After the integration, users of SCF were required to choose among CCN’s categories: ‘damaged sign’, ‘graffiti’, ‘other’, ‘pothole’, ‘sidewalk patch’, ‘street light’, and during the winter months: ‘roadway plowing and sanding’, and ‘unshoveled sidewalk’. This event created a unique opportunity to observe how the introduction of a category system affected reporting behavior, with the interface, location, and user group remaining largely the same. While the overall number of reports submitted to SCF is small (only counting initial reports without comments), an increase in Graffiti-related reports can be observed after the introduction of the CCN categories: from 2 in 167 (1.2%) reports submitted until Sept. 2011 to 11 in 255 (4.3%) reports submitted since Sept. 2011. More than two thirds (75%) of SCF reports were submitted to the ‘other’ category, compared to 52% of the reports submitted to CCN during the same period, which is consistent with the earlier observation that users of SCF tend to use the system for discussing wider civic topics rather than service requests (Offenhuber, 2014).

Dissecting the ‘other’ category

A second aspect of the relationship between offered categories and reporting behavior concerns the question what kinds of incidents are submitted to the ‘other’ category, or put differently, which types of incidents are masked by the ‘other’ category. As of June 2014, more than 34800 reports have been submitted under the ‘other’ category to CCN since its launch in fall 2010, which provides a rich corpus of documents for further text analysis. While manual content analysis was used to extract subtleties of individual reports such as tone, motivations or styles of justification, the manual approach reaches its limits dealing with a large body of documents in order to discover most salient themes.

For this purpose, I chose a probabilistic topic modeling approach, using the ‘Mallet’ and R software toolkits. After training topic models for 10 to 40 topics in steps of five, I chose 30 topics over 500 iterations, based on considerations of topic size and specificity. The topics returned by the algorithm have to be taken with a grain of salt – not every topic represents a meaningful theme; meaningful themes may span multiple

---

12 [http://mallet.cs.umass.edu/topics.php](http://mallet.cs.umass.edu/topics.php)
13 [http://www.r-project.org/](http://www.r-project.org/)
topics or conversely, a single topic can contain multiple unrelated themes (Schmidt, 2013). The interpretation of topics therefore requires an additional review of reports that are highly associated with a specific topic. I have chosen a probability value of $p=.85$ as the threshold for associating a report with a topic.

Before running the estimation, stop-words have been removed from the corpus (i.e. ‘but’, ‘this’, ‘the’), including a custom list of stop-words containing street names, pre- and suffixes, directions (i.e. north, south, east, west) as well as person names. Stemming was not used, as it tends to produce ambiguous results in this context (i.e. ‘parking’ and ‘park’ would become one word), at the expense of having sometimes the same word multiple times in a single topic (i.e. ‘pothole’ and ‘potholes’).

Figure 6 Selected topics extracted from reports submitted to the CCN-Other category since Oct. 2010. The y axis represents the logarithmic normalized frequency of the term within all reports; the x axis represents the level of specificity of the term for the respective topic. For a full list of topics refer to the Appendix.
Figure 6 shows a selection of topics detected within the ‘other’ category, sorted by the number of associated documents. The full list of topics can be found in the appendix (Figure 7). The most frequently identified topic 16 concerns issues of illegally parked cars obstructing and blocking traffic. Example reports include “Bus blocking fire Hydrant and street. Issue not resolved”, or “Driver parked several feet off curb, obstructing traffic”. Other topics associated with parking violations are 1 and 8, which both contain reports about vehicles parking without a resident permit or cases of neighbors trying to reserve their parking spot with cones other means: “372Lrg FL TAG PARKING NO STICKER doing this for week and not getting a ticket” and “No place to park and funeral home is coning off public street and getting mad at residents”. The second ranked topic (30) in terms of number of associated documents concerns issues of garbage bags left on the sidewalk “394 Riverway has dumped many bags and loss bit of trash on street. This is not a trash pickup location and it is not trash day. This is a frequent issue with this address please cite them. Mission Hill is not a dump”. Other garbage related topics include overflowing public waste bins (topic 17) and rats and rodents attracted by trash (topic 10): “Neighbor across the street is continuously dumping rice and other food here, attracting rats, mice, and other pests”. Dangerous traffic situations are also frequently reported (topics 14, 15), as well as issues of fallen trees and branches (topic 26) and overgrown weed (topic 3). Less frequent, but salient examples include noise complaints (topics 19), such as: “huge backyard student party at 806 parker st. VERY loud, underage?”), and complaints about dog owners and homeless people (topic 25). Topic 29 represents a group of consistent reports submitted by locals in the North End who use CCN to advocate and protest against the a Segway tours operator, which they perceive as a nuisance: “Illegal Boston Gliders. Boston by Segway tour monopolizing Long Wharf pedestrian park”. Topic 13 represents complaints about a previously closed issue that is perceived as still being unresolved and difficulties with the CCN app: “Getting close to one month with no update and no action. Is this how this system is supposed to work?”

To facilitate comparison, I trained a topic model also the whole corpus of CCN reports covering all categories including the general category (Figure 8). The extracted topics reveal the standard service categories including potholes (topic 23), graffiti (topic
15, 22), side walk repair (topic 17), streetlights (topic 6), snow issues (topic 9) and damaged signs (topic 14). However, in terms of the number of associated documents, the issues identified in the previous model still dominate the identified topics. This might be related to two different reasons. First, many of the pre-defined service categories may be under-represented because they do not require lengthy textual explanations, such as in the case of graffiti and potholes, while ‘other’ issues nearly always require a description of the issue. Second, the ‘other’ category is by far the largest category submitted, and it should not be surprising that some of the themes within this category are more prevalent than some of the specific categories. For example, trash and parking issues within ‘other’ are more frequently reported than for example “damaged sign” issues. This leads to the interesting question, why the city would offer a category for damaged signs but not for the more frequent parking violations or garbage issues?

**Discussion - a case for ambiguity**

In this process of designating a set of service categories, a city can choose a number of different options, or a combination thereof:

1. **Should the categories correspond to internal structures or the expectations of the user?** In general, interfaces should be always user-centered, and correspond to how users see the system (Norman, 2002). Yet, it is unavoidable to think about the mapping to internal structures to ensure that issues are not only received, but also resolved.

2. **Should categories correspond to departmental responsibilities?** A number of cities have chosen this route by offering categories such as ‘police issue’, ‘sanitation issue’ and so forth. If responsibilities are not obvious from a report, the city can use such reports to delineate boundaries between the different departments and develop protocols for resolving such issues.

3. **Should categories emerge from user requests without offering initial categories at all?** Known from social media as a folksonomy approach (Voss, 2007), users would tag their reports with keywords they can freely choose. As more and more
reports are submitted, frequent keywords are suggested by the system, which reinforces convergence.

4. Should categories be selected based on their actionability? The downside of a folksonomy approach might be that emerging categories or keywords might not necessarily the most actionable one for the city. A complaint that a park is neglected and should be improved is difficult to address. On the other hand, a report about a damaged sign or a pothole can immediately be added to the service queue.

Considering all these issues, every solution requires an iterative process negotiating user perceptions, departmental boundaries, and technical constraints of the interface. In this process, ambiguities cannot be avoided and should be considered as a resource, not a problem (Gaver, Beaver, & Benford, 2003). For example, users are not necessarily aware of exact shape of city boundaries, local, state and federal responsibilities, whether the issue concerns public or private property or that fire or police departments are considered outside the realm of 311 systems. As Nigel Jacob from the Boston office of New Urban Mechanics argues, reporting applications should nevertheless be designed to allow such cases beyond the city’s responsibility and use these reports to develop interfaces with outside entities. The ‘other’ category assumes an important role to capture these ambiguities, and separate them for incidents that can be more readily addressed.

A second aspect that is less clear-cut than often thought is the distinction between citizen and city employee. The CitizensConnect app is not only popular with residents, but also with city workers, who appreciate the simplicity of submitting incidents, as revealed by the reports and confirmed by the Office for New Urban Mechanics. To respond to that need, the office also offers a version of the smartphone app for city employees.14 Similarly, SeeClickFix represents all users, whether they are residents, representatives or individual city employees in much the same way, which encourages interaction at eye-level. Unlike in CCN, users of SCF also have the capacity to reopen issues if they consider them not being resolved.

14 link to city worker app
Conclusion

Categories are just one way how the design of civic technologies such as citizen feedback applications shapes the usage and consequently the data collected through these interfaces. Other aspects include the management of visibility of the reports, the social presence and representation of other users, the affordances of the interface and many more. A close inspection of any of these design patterns reveals that even mundane design details leave their imprint on the collected data, which cannot be accurately captured on a one-dimensional truth/bias spectrum. If the analysis of large data sets is limited to the data itself and does not consider the way it was collected, much information is lost.

The second conclusion is that designers of civic technologies are not only in charge of the user experience, their work has also impact on governance by regulating and framing the interaction between citizens and the city. The setup of a traditional urban planning meetings involving citizen participation is recognized as a highly political issue. In the case of digital participation, it is generally overlooked that the design of the interface plays a very similar role (Galloway, 2012).

This paper makes a case for critically scrutinizing interfaces and their effects on civic participation and data collection. Furthermore, it also highlights the potential for a new form of urban governance, as civic technologies and participatory platforms increasingly enter the planning domain. The political nature of interface shows that the definition of standards, protocols or applications should be also considered a critical component of democratic discourse.

References


### Appendix

<table>
<thead>
<tr>
<th>Topic</th>
<th>topic label</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>parked parking cars</td>
<td>311</td>
</tr>
<tr>
<td>30</td>
<td>trash garbage street</td>
<td>294</td>
</tr>
<tr>
<td>1</td>
<td>parked car resident</td>
<td>253</td>
</tr>
<tr>
<td>18</td>
<td>trash pick up</td>
<td>235</td>
</tr>
<tr>
<td>26</td>
<td>tree dead trees</td>
<td>146</td>
</tr>
<tr>
<td>14</td>
<td>traffic light lights</td>
<td>140</td>
</tr>
<tr>
<td>21</td>
<td>trash property lot</td>
<td>118</td>
</tr>
<tr>
<td>8</td>
<td>parking space street</td>
<td>112</td>
</tr>
<tr>
<td>15</td>
<td>stop sign crosswalk</td>
<td>107</td>
</tr>
<tr>
<td>22</td>
<td>street lane turn</td>
<td>90</td>
</tr>
<tr>
<td>9</td>
<td>walk signal pedestrian</td>
<td>88</td>
</tr>
<tr>
<td>11</td>
<td>fence missing broken</td>
<td>85</td>
</tr>
<tr>
<td>20</td>
<td>sidewalk blocking street</td>
<td>81</td>
</tr>
<tr>
<td>5</td>
<td>park playground broken</td>
<td>68</td>
</tr>
<tr>
<td>17</td>
<td>trash big overflowing</td>
<td>66</td>
</tr>
<tr>
<td>10</td>
<td>rats trash rat</td>
<td>57</td>
</tr>
<tr>
<td>27</td>
<td>bike remove abandoned</td>
<td>57</td>
</tr>
<tr>
<td>6</td>
<td>sidewalk left remove</td>
<td>54</td>
</tr>
<tr>
<td>12</td>
<td>water cover sidewalk</td>
<td>52</td>
</tr>
<tr>
<td>4</td>
<td>drain water storm</td>
<td>51</td>
</tr>
<tr>
<td>7</td>
<td>construction permit signs</td>
<td>51</td>
</tr>
<tr>
<td>19</td>
<td>night loud noise</td>
<td>51</td>
</tr>
<tr>
<td>24</td>
<td>street cleaning side</td>
<td>45</td>
</tr>
<tr>
<td>28</td>
<td>box pole wires</td>
<td>41</td>
</tr>
<tr>
<td>2</td>
<td>broken glass sidewalk</td>
<td>39</td>
</tr>
<tr>
<td>25</td>
<td>park dog people</td>
<td>39</td>
</tr>
<tr>
<td>29</td>
<td>illegal park segway</td>
<td>35</td>
</tr>
<tr>
<td>23</td>
<td>city people time</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>grass weeds overgrown</td>
<td>23</td>
</tr>
<tr>
<td>13</td>
<td>reported closed case</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unique word across topics</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
</tr>
<tr>
<td>no</td>
</tr>
</tbody>
</table>

Figure 7 Topics extracted from all reports submitted to CCN-other category since October 2010, ordered by number of associated reports with a probability > 0.85. Blue words are unique to the topic; grey words also appear in other topics. Topic labels are the three most salient words for the specific topic.
Figure 8 Topics extracted from all reports submitted to CCN since October 2010, ordered by number of associated reports with a probability > 0.85. Blue words are unique to the topic; grey words also appear in other topics. Topic labels are the three most salient words for the specific topic.
How well do we understand US neighborhood? The Potential for Big Data to Improve Neighborhood-Level Census Data

Seth E. Spielman
Assistant Professor of Geography, University of Colorado.
seth.spielman@colorado.edu

Abstract
The promise of “big data” for those who study cities is that it offers new ways of understanding urban environments. The potential of big data is, at least in part, constrained by the broader national data economies within which big data exist, this data economy has changed in ways that are both poorly understood by the average data consumer and of consequence to the application of big data to urban problems. For example, high resolution demographic and economic data from the Census Bureau are terribly imprecise, for some policy-relevant variables, like the number of children under 5 in poverty, the estimates are almost unusable, of the 56,204 tracts for which a poverty estimate available 40,941 had a margin of error greater than the estimate in the 2007-2011 ACS (72.8% of tracts). For example, the ACS indicates that Census Tract 196 in Brooklyn, NY has 169 children under 5 in poverty ± 174 children, suggesting somewhere between 0 and 343 children in the area live in poverty. While big data is exciting and novel, basic questions about American Cities are all but unanswerable in the current data economy. In this paper we describe how big data might be used improve tract-level census data through intervention in four specific areas (see bold text in the manuscript for a summer of these areas).
**Introduction**

The promise of “big data” for those who study cities is that it offers new ways of understanding urban environments and/or their affect on human behavior. Big data lets one see urban dynamics at much higher spatial and temporal resolutions than the more traditional sources of data, such as the survey data typically collected by public agencies. Some see the rise of big data as a revolutionary mode of understanding cities, this “revolution” is which is enticing to academics because, as argued by Kitchin (2014), revolutions in science are often preceded by revolutions in measurement. That is, big data could give rise to something even bigger, a new science of cities. Others, such as Greenfield (2014) argue that real urban problems cannot be solved by data and are deeply skeptical of the potential for information technologies to have meaningful impacts on urban life. Here, we aim to contextualize big data within a broader national data economy. The emergence of big data is not the only recent “revolution” in the measurement of urban environments. The potential of big data is, at least in part, contingent upon the context within which big data exists, this context has changed in ways that are both poorly understood by the average data consumer and of consequence to the application of big data to urban problems.

Traditional sources of information about cities have recently changed in profound ways, we argue that these changes create potential, and problems, for the application of big data to urban questions.

The data collected by public agencies about American cities has recently undergone a series of dramatic changes, some of these changes are a result of the gradual accrual of broader social changes and some have been abrupt, the result of changes to federal policy. Groves et al. (2013) document a gradual long term national trend of increases in the number of people who refuse to respond to public (and private) surveys. Geographic and demographic patterns in survey non-response make it difficult for surveys to accurately describe populations and create the need for complex statistical adjustments to ensure that the statistics produced by the survey are representative of the target population. If for example, low income, non-documented immigrants do not respond to official surveys they would be invisible to the data-centric urban analyst. If

---

1 Defining big data is difficult, most existing definitions, include some multiple of V’s (see Laney 2001). All are satisfactory for our purposes here. We use the term to distinguish between census/survey data which we see as “designed” measurement instruments and big data which we see as “accidental” measurement instruments.
undocumented immigrants respond only rarely, any assessment of their prevalence would be subject to enormous uncertainty.

In fact, high levels of uncertainty now plagues almost all fine resolution urban data produced by the United States Census Bureau. Neighborhood-level data from the Census Bureau are terribly imprecise, for some policy-relevant variables, like the number of children in poverty, the estimates are almost unusable — of the 56,204 tracts for which a poverty estimate for children under 5 was available 40,941 had a margin of error greater than the estimate in the 2007-2011 ACS (72.8% of tracts). For example, the ACS indicates that Census Tract 196 in Brooklyn, NY has 169 children under 5 in poverty ± 174 children, suggesting somewhere between 0 and 343 children in the area live in poverty. Users of users often face data like those in Table 1, which shows the ACS median income estimates for African American households for a contiguous group of census tracts in Denver, Colorado. Income estimates range from around $21,000 to $60,000 (American Factfinder website accessed 7/15/2013). Without taking account of the margin of error, it would seem that Tract 41.06 had the highest income, however, when one accounts for the margin of error, the situation is much less clear – Tract 41.06 may be either the wealthiest or the poorest tract in the group.

<table>
<thead>
<tr>
<th>Tract Number</th>
<th>African-American Median Household Income</th>
<th>Margin of Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Census Tract 41.01</td>
<td>$28,864</td>
<td>$8,650</td>
</tr>
<tr>
<td>Census Tract 41.02</td>
<td>$21,021</td>
<td>$4,458</td>
</tr>
<tr>
<td>Census Tract 41.03</td>
<td>$43,021</td>
<td>$14,612</td>
</tr>
<tr>
<td>Census Tract 41.04</td>
<td>$36,092</td>
<td>$3,685</td>
</tr>
<tr>
<td>Census Tract 41.06</td>
<td>$60,592</td>
<td>$68,846</td>
</tr>
</tbody>
</table>

Table 1: 2006-2010 ACS Estimates of African-American Median Household Income in a selected group of proximal tracts in Denver County, Colorado

---

2 We use the terms “fine” and “high” resolution to refer to census tract or smaller geographies, these data are commonly conceived of as “neighborhood-scale” data. We conceive of resolution in the spatial sense, higher/finer resolution means a smaller census tabulation unit. However, the geographic scale high resolution of census units is a function of population density.
The uncertainty in table 1 is all but ignored by practicing planners. We recently conducted a survey of 180 urban planners and found that most planners (67%) simply delete or ignore the rightmost column when preparing maps and reports. This practice, according to planners, is driven by the “demands” of their “consumers.” That is, the audience for their maps and reports would have difficulty incorporating the margins of error into decision making processes. This practice is further reinforced by federal agencies, which use the tract level estimates only to determine eligibility for certain programs (for example, Treasury’s New Markets Tax Credit program). The problem with the margins of error is especially pronounced for the Census Transportation Planning Package, a key into transportation planning processes.

The decline in the quality of neighborhood scale data in the United States began in 2010, the year the American Community Survey (ACS) replaced the long form of the United States decennial census as the principal source of high-resolution geographic information about the U.S. population. The ACS fundamentally changed the way data about American communities are collected and produced. The long form of the decennial census was a large-sample, low-frequency national survey; the ACS is a high-frequency survey, constantly measuring the American population using small monthly samples. One of the primary challenges for users of the ACS is that the margins of error are on average 75 percent larger than those of the corresponding 2000 long-form estimate (Alexander 2002; Starinic 2005).

Some degree of uncertainty is inherent in surveys like the ACS, however the amount of uncertainty in the ACS has far exceeded the United States Census Bureau’s (USCB hereinafter) expectations. Initial expectations were that the amount of uncertainty (margin of error) in the ACS would be 33 percent greater than the decennial census long form (Navarro 2012). This loss in precision was justified by the increase in timeliness of ACS estimates which are released annually compared to the once a decade long form. This tradeoff prompted Macdonald (2006) to call the ACS a “warm” (current) but “fuzzy” (imprecise) source of data. Unfortunately, those early expectations were too optimistic, the actual uncertainty in the ACS is much more than 33 percent greater than the census long form. While there are clear advantages to working with fresh
data, the ACS margins of error are so large that for many variables at the census tract and block group scales the estimates fail to meet even the loosest standards of data quality.

Many of the problems of the American Community Survey are rooted in data limitations. That is at critical stages in the creation of neighborhood-level estimates the census bureau lacks sufficient information and has to make assumptions and/or use data from a coarser level of aggregation (municipality or county). We argue that one of the major potential impacts of big data for the study of cities is the reduction of variance in more traditional forms demographic and economic information. TO support this claim we describe the construction of the ACS in some detail, with the hope that this details illuminates the potential for big data to improve federal and/or state statistical programs.

Understanding the American Community Survey
Like the decennial long form before it, the ACS is a sample survey. Unlike complete enumerations, sample surveys do not perfectly measure the characteristics of the population—two samples from the same population will yield different estimates. This sample-to-sample variability creates some uncertainty about a population’s true characteristics, therefore survey-based estimates are usually accompanied by a margin of error. While it was not commonly acknowledged, even the decennial census long form data came with instructions for estimating margins of error. In the ACS, the margin of error for a given variable expresses a range of values around the estimate within which the true value is expected to lie. The margin of error reflects the variability that could be expected if the survey were repeated with a different random sample of the same population. This variability is referred to as sampling error and is measured as standard error (SE). Calculating standard errors for a complex survey like the ACS is not a trivial task, the USCB uses a simulation procedure called Successive Differences Replication to produce variance estimates (Wolter, 1984; Fay & Train, 1995; Judkins, 1990). The margins of error reported by the USCB with the ACS estimates are simply 1.645 times the standard errors.

Sampling error has two main causes. The first is the sample size - the larger the sample the smaller the standard error, intuitively more data about a population leads to less uncertainty about its true characteristics. The second main cause of sampling error is heterogeneity in the
population being measured (Rao 2003). Consider two jars of U.S. coins, one contains U.S. pennies and the other contains a variety of coins from all over the world. If one randomly selected 5 coins from each jar, and used the average of these 5 to estimate the average value of the coins in each jar, then there would be more uncertainty about the average value in the jar that contained a diverse mixture of coins. If one took repeated random samples of 5 coins from each jar the result would always be the same for the jar of pennies but it would vary substantially in the diverse jar, this variation would create uncertainty about the true average value. While the ACS is much more complicated than pulling coins from a jar, this analogy helps to understand the standard error of ACS estimates. Census Tracts and block groups are like jars of coins. If a tract is like the jar of pennies, than the estimates will be more precise, whereas if a tract is like the jar of diverse coins, then the estimate will be less precise.

While the simple example is illustrative of important concepts it overlooks the central challenge in conducting surveys; many people who will be included in a sample will choose not to respond to the survey. While a group’s odds of being included in the ACS sample are proportional to its population size, different groups of people have different probabilities of responding. Only 65% of the people contacted by the ACS actually complete the survey (in 2011, 2.13 million responses were collected from 3.27 million samples). Some groups are more likely to respond than others, this means that a response collected from a hard to count group is worth more than a response from an easy to count group. These differential response rates are controlled by weighting each response. In the ACS each completed survey is assigned a single weight through a complex procedure involving dozens of steps. The important point, as far as this paper is concerned, is that these weights are estimated and uncertainty about the appropriate weight to give each response is an important source of uncertainty in the published data.

The concepts of sampling error and weighting are central to understanding uncertainty in the ACS but they are not the entire story. Some of the factors affecting uncertainty in the ACS are

---

3 The Census Bureau generally is not actually estimating the “average” value, they are estimating the “total” value of coins in the jar. Repeatedly grabbing five coins and computing the average will over many samples get you a very precise estimate of the average value, but it will give you no information on the total value. To get the total value, you need a good estimate of the average AND a good estimate of the total number of coins in the jar. The loss of cotemporaneous population controls caused by decoupling the ACS from the Decennial enumeration means that the census does not have information about the number of coins in the jar. This is discussed in more details later.
the result of decisions and tradeoffs made by the USCB, whereas other factors affecting uncertainty are the result of circumstances beyond the control of the USCB.

The data-quality problems in the ACS are the direct result of decades of innovation in national surveys. Four innovations in particular are combined in the ACS:

1. The use of sampling
2. The provision of small-area estimates.
3. The release of annual estimates.
4. The use of complex model-based weighting procedures to adjust the importance of individual responses.

Each of these innovations is constrained by the availability of data. This is where big data might help…

**Sampling**

Before 1940, there was no “long form” or “short form” of the U.S. decennial census; each housing unit (HU) received the same questionnaire. By 1940 the census forms had become long and complicated, through the gradual accrual of demographic and economic questions. In response, the questionnaire was split in 1940 into a short set of questions asked of 100 percent of the population and an additional “long form” administered to a subset of the population. Originally, this long form was administered to a 5 percent random sample, but in later years it was sent to one HU in six (Andersen et al. 2011). Before 1940 any error in the data could be attributed either to missing or double counting a HU, to incorrect transcription of a respondent’s answer, or to intentional/unintentional errors by the respondent. After 1940, however, the adoption of statistical sampling introduced new sources of uncertainty for those questions on the long form. The bifurcation into short and long forms continued through the 2000 census, but in 2010 the decennial census returned to a single, short questionnaire. The detailed demographic, economic, and housing data previously collected on the long form are now provided by the ACS.

The American community Survey constantly measures the population, it does not co-occur with a complete census. The lack of complete count population data is a key source of uncertainty in the ACS. Prior to the rise of the ACS, short form population counts could serve as controls for
long-form based estimates. The decoupling of the sample from the complete enumeration accounts for 15 to 25 percent of the difference in margin of error between the ACS and the decennial long form (Navarro 2012). Population controls are essential to the ACS sample weighting process, now population controls are only available for relatively large geographic areas such as municipalities and counties.

If big data could be used to produce reliable, timely, population counts for census blocks or tracts, it would be possible to substantially reduce the margin of error in more detailed demographic and economic estimates.

Small Area Estimates
The geographic resolution of ACS estimates also contributes to uncertainty. Tract-level census tabulations were first proposed by Walter Laidlaw, a Presbyterian minister, in 1906. The New York Times in 1923, clearly excited by the potential of Laidlaw’s tract system, noted that it allowed one to “know precisely and for the first time what is meant by New York.” While survey methodology has evolved substantially over the past century, the practice of reporting estimates for small geographic units has remained fundamentally unchanged. The definitions of some geographic units have changed, and new types of geographic units have been created, but the basic idea of tabulating results using relatively static geographic zones remains constant. The principle is that tract (and smaller) geographic units are designed using criteria that are largely exogenous to the detailed demographic data collected by the census. The ACS and the decennial census use the same system of blocks, block groups, and census tracts for reporting estimates.

Prior to the advent of sampling, the complete count census data could, in principle, be tabulated using any sort of geographic zone. The advent of sampling made small area census data less precise. Since there are a finite number of samples in any geographic area, as tabulation zones become smaller sample sizes decline, making the ACS estimates more uncertain. The rise in uncertainty is greater for small populations; for instance the effects of reducing a sample size from 200 to 100 is much greater than the effect of reducing a sample size from 20,000 to 10,000. The USCB counteracts this decline in sample size by pooling small area samples over multiple years, thus diluting the temporal resolution of the estimates. Larger areas do not require
multiyear pooling. For large municipalities, counties and states the ACS is a distinct improvement over the decennial census because it provides high quality annual data, however this may not be true at the tract/block group scale where even the 5 year pooled data have much larger standard errors and smaller sample sizes than the 2000 long form data.

The design of tabulation geographies could be data driven (rather than manual). Moreover, thinking back to the jars of coins example, if these geographies contained a more homogenous population it might be possible to improve survey estimates. Moreover, these geographies would need to find the appropriate balance between geographic detail and data quality. Big data provides an entirely new basis for the segmentation of urban space:

**Big data might allow the data-driven design of tabulation units that are rooted in human behavior, for example using circulation and activity patterns to segment urban space**\(^4\). These units would have the advantage of having a meaningful interpretation and more precise statistical descriptions.

Annual Estimates

Small-area tabulations and the survey-based long form had become cornerstones of the decennial census by the late twentieth century. However, data users were increasingly concerned about the timeliness of the once-a-decade data (Alexander 2002). In 1985 Congress authorized, but never funded, a mid-decade census to address this problem. Throughout the 1980s and early 1990s interest in a “continuous measurement” model evolved within the USCB, leading to the proposal of the Intercensal Long Form (Alexander 1993). Continuous measurement was inspired by Leslie Kish, a statistician who developed the theory and methods for rolling surveys (Kish 1990).

Kish’s basic idea was that a population could be divided into a series of non-overlapping annual or monthly groups called subframes. Each subframe would then be enumerated or sampled on a rolling basis. If each subframe were carefully constructed so as to be representative of the larger population, then the annual estimates would also be representative, and eventually, the entire

\(^4\) This may not be possible… Herbert Simon, in the *Architecture of Complexity* (1961) suggests that complex systems, like a cities, should be divided into units of analysis on the basis of information flow (not power relations). The idea is that units should have more internal than external interaction/exchange. While one might be able to delineate an individuals activity space, it is not clear if these would be generalizable. It seems like an interesting research question, are there geographically bounded urban units (such as neighborhoods) that meet Simon’s criteria?
population would be sampled. The strength of this rolling framework is its efficient use of surveys. The decennial census long form had to sample at a rate appropriate to make reasonable estimates for small geographic areas such as census tracts, which contain on average 4,000 people. Therefore, citywide data released for a municipality of, say, 1 million people would be based on considerably more samples than necessary. Spreading the samples over time lets larger areas receive reasonable estimates annually, while smaller areas wait for more surveys to be collected. The rolling sample therefore increases the frequency of data on larger areas. The primary cost comes in the temporal blurring of data for smaller areas.

Rolling sampling is straightforward in the abstract. For example, suppose that there are K=5 annual subframes, that the population in a tract is known (N=1000), that the sampling rate is r=1/6, and that the response rate is 100 percent; then one would sample \( n = \frac{N}{K*1/r} \) people per year. Over a 5 year period 1/6 of the population would be sampled and each returned survey would represent \( w = \frac{N/n}{K} \) people, where \( w \) is the weight used to scale survey responses up to a population estimate. In this simple case, the weight assigned to each survey would be the same. For any individual attribute \( y \), the tract level estimate would be \( y_t = \sum w_i y_i \) (equation 1), a weighted summation of all \( i \) surveys collected in tract \( t \). If the weights are further adjusted by ancillary population controls \( X \), then the variance of the estimate is \( \Sigma w_i^2 \text{VAR}[y_i|X] \) (equation 2; Fuller 2011, assuming independence.). If the rolling sample consisting of long-form-type questions were administered simultaneously with a short form census, then all the parameters in our simple example (\( N, K, X \)) would be known.

However, in the ACS good population controls are not available for small areas (\( N \) and \( X \) are unknown) because, unlike the long form, the survey is not contemporaneous with the complete enumeration decennial census. Thus weights (\( w \)) for each response must be estimated and this is an important source of uncertainty in the ACS. The lack of such controls accounts 15-25% of the increase in variance between the decennial long form and the ACS (Navarro 2012).

**Big data might be useful for establishing population controls to allow more effective estimation of survey weights.**
Weighting

In the ACS each completed survey is assigned a weight \( w \) that quantifies the number of persons in the total population that are represented by a sampled household/individual. For example, a survey completed by an Asian male earning $45,000 per year and assigned a weight of 50 would in the final tract-level estimates represent 50 Asian men and $2.25 million in aggregate income. The USCB’s careful attention to survey weights ensures that the final tract-level estimates are as accurate as possible, i.e., unbiased. Variations in sampling rate, the lack of good estimates of tract-level population controls, and variations in response rate all necessitate a complex method to estimate \( w \).

The construction of ACS weights is described in the ACS technical manual (which runs hundreds of pages, USCB 2009a). The complexity of the ACS weighting process is motivated by an effort to reduce bias, each step can be seen as an attempt to control some form of bias. For example, the temporal adjustments applied to weights are an attempt to control for bias that might arise due to seasonal fluctuations in population, “mode bias” adjustments are made to control for differences in surveys completed by mail, phone, internet, and in-person interview. Individually these steps make sense but they are so numerous and technically complex that in the aggregate they make the ACS estimation process nearly impenetrable for even the most sophisticated data users.

The cost of extensive tweaking of weights is more than just lack of transparency and complexity. Reducing bias by adjusting weights carries a cost. Any procedure that increases the variability in the survey weights also increases the uncertainty in tract-level estimates (Kish 2002). Embedded in this process is a trade-off between estimate accuracy (bias) and precision (variance/margin of error), refining the survey weights reduces bias in the ACS but it also leads to variance in the sample weights. Increasing the variation in the sample weights increases the margin of error of the final estimates (Alexander 2002; Kish 2002) (see equation 2). The weighting procedures applied by the USCB in the estimation of the ACS can then be seen as an implicit policy statement that unbiased (accurate) estimates are more important than precise (low-variance) estimates. While weighting procedures are not solely responsible for the low precision in ACS small-area estimates, the current bias-versus-variance calculation may need to be reevaluated.
There are other census programs that publish small area estimates with lower variance but with higher bias, one such example is the Small Area Income and Poverty Estimates (SAIPE) program. It is not clear from the published material about the ACS how much each weighting step reduces bias and increases variance.

Part of the reason it is difficult to estimate survey weights is that there is little information available about survey respondents prior to their selection into the sample. The USCB uses a method called Generalized Regression (GREG) to reduce uncertainty in small-area estimates by reducing the variance of the sample weights. The ACS GREG procedure incorporates person-level administrative data from federal agencies on age, race, and gender. Estimates for large counties and cities do not use GREG. In order to maintain a single methodology for the entire country, the GREG procedure only uses variables with national coverage.

**Big Data could provide more detailed individual/household level information, to allow the more effective estimation/refinement of survey weights.**

The limited set of variables currently used could potentially be expanded to include real estate sales or other transactional databases. For example, home prices exhibit strong spatial autocorrelation and therefore it might be possible to use a geostatistical framework to estimate home prices - an important aspect of a household’s economic characteristics. Through the currently used GREG procedure such a database would have the potential to improve the ACS weights. Research into both of these strategies - of defining new geographic units through aggregation, and improving spatial data-fusion techniques to improve the estimation ACS weights - are areas in which geographers can play a unique role.

**Conclusion**

Little (2012) argues that a fundamental philosophical shift is necessary within both federal statistical agencies and among data users, “we should see the traditional survey as one of an array of data sources, including administrative records, and other information gleaned from cyberspace. Tying this information together to yield cost-effective and reliable estimates...” However, Little also notes that for the Census “combining information from a variety of data
sources is attractive in principle, but difficult in practice” (Little 2012, p.309). By understanding the causes of uncertainty in the ACS the implications of Little’s statement become clear, there is enormous potential to mash-up multiple forms of information to provide a more detailed picture of US cities.

References


Big Data and Survey Research: Supplement or Substitute?

Timothy P. Johnson
Survey Research Laboratory
University of Illinois at Chicago
412 S. Peoria St.
Chicago, IL 60607
timj@uic.edu

and

Tom W. Smith
General Social Survey
NORC at the University of Chicago
1155 E 60th St.
Chicago, IL 60637
Smith-tom@norc.org

To be presented at the Workshop on Big Data and Urban Informatics, University of Illinois at Chicago, August 2014.
Abstract

The increasing availability of organic Big Data has prompted questions regarding its usefulness as an auxiliary data source that can enhance the value of design-based survey data, or possibly serve as a replacement for it. Big Data’s potential value as a substitute for survey data is largely driven by recognition of the potential cost savings associated with a transition from reliance on expensive and often slow-to-complete survey data collection to reliance on far less-costly and readily available Big Data sources. There may be, of course, methodological costs of doing so. We review and compare the advantages and disadvantages of survey research and Big Data methodologies, concluding that each data source has unique qualities and that future efforts to find ways of integrating data obtained from varying sources, including Big Data and survey research, are most likely to be fruitful.
Introduction

As response rates and survey participation continue to decline, and as costs of data collection continue to grow, researchers are increasingly looking for alternatives to traditional survey research methods for the collection of social science information. One approach has involved modifying scientific survey research methods through the abandonment of probability sampling techniques in favor of less expensive non-probability sampling methodologies (c.f. Cohn, 2014). This strategy has become popular enough that the American Association for Public Opinion Research (AAPOR) has recently felt it necessary to appoint a Task Force to investigate the issue and release a formal report (Baker et al., 2013). Others have explored the usefulness of supplementing, or replacing completely, surveys with information captured efficiently and inexpensively via “Big Data” electronic information systems. In this paper, we explore the advantages and disadvantages of using survey data versus Big Data for purposes of social monitoring and address the degree to which Big Data might become a supplement to survey research or a complete alternative or replacement for it.

Survey research evolved out of social and political needs for better understandings of human populations and social conditions (Converse 1987). Its genesis predates considerably the pre-electronic era when there were few alternative sources of systematically collected information. Over the past 80 years, survey research has grown and diversified, and complex modern societies have come to increasingly rely on survey statistics for a variety of public and private purposes, including public administration and urban planning, consumer and market research, and academic investigations, to name a few. In contrast, Big Data became possible only recently with the advent of reliable, high speed and relatively inexpensive electronic systems capable of prospectively recording vast amounts of seemingly mundane process information. In a very short period of time, Big Data has demonstrated its potential value as an alternative method of social analysis (Goel et al., 2010; Mayer-Schönberger and Cukier, 2013).

Before proceeding further, however, it is important to define what we mean exactly by survey research and “Big Data.” Vogt (1999: 286) defines a survey as “a research design in which a sample of subjects is drawn from a population and studied (often interviewed) to make inferences about the population.” Groves (2011) classifies surveys as forms of inquiry that are “design-based,” as the specific methodology implemented for any given study is tailored (or designed) specifically to address research questions or problems of interest. In contrast, Webpodeia (2014) defines Big Data as “a buzzword…used to describe a massive volume of both structured and unstructured data that is so large that it’s difficult to process using traditional database and software techniques.” In addition to these attributes, Couper (2013) observes that Big Data is produced at a rapid pace. In contrast to design-based data, Groves classifies Big Data as being organic in nature. Although similar to survey data in the systematic manner in which it is collected, organic data is not typically designed to address specific research questions. Rather, such data, referred to by Harford (2014) as “digital exhaust,” is a by-product of automated processes that can be quantified and reused for other purposes. There are, of course, exceptions, such as the National Weather Service’s measurements, which are design-based and otherwise fit the definition of Big Data.

Although they do not fit today’s electronic-based definitions of Big Data, there are several examples of survey-based data sets that are uncharacteristically “big” by any reasonable standards. Examples of Big Surveys include national censuses, which routinely attempt to collect
information from millions of citizens. The U.S. micro decennial Census is an example of this. Also included here is the infamous *Literary Digest* Poll, which attempted, and failed badly, to predict the outcome of the 1936 Presidential election, based on more than two million postcard responses collected from individuals sampled from published telephone directories and automobile registration lists (Squire, 1988). The *Literary Digest* had been conducting similar straw polls since 1908, but did not run into trouble with a failed election prediction until 1936. The *Literary Digest* experience taught the still young survey research community of the 1930s that big does not necessarily mean better. Subsequent to that experience, survey statisticians worked to develop sampling theory, which enabled them to rely on much smaller, but more carefully selected, random samples to represent populations of interest.

**What Distinguishes Surveys from Big Data?**

While censuses and the *Literary Digest* examples share with today’s Big Data large observation-to-variable ratios, they do not have Big Data’s electronic-based longitudinal velocity, or rate of data accumulation. Rather, even Big Surveys are only snapshots that represent at best a brief moment in time. Perhaps even more importantly, the structures of these design-based data sources are carefully constructed, unlike many sources of Big Data, which are known for their “messy” nature (Couper, 2013). Hence, there are several important differences between design-based survey data, and the organic data sources that represent Big Data. These include differences in volume, data structures, the velocity and chronicity with which data are accumulated, and the intended purposes for which the data are collected.

**Volume**

Big Data is big by definition. As Webpodeia (2014) suggests, Big Data represents “a massive volume of both structured and unstructured data that is so large that it’s difficult to process using traditional database and software techniques.” Most of the data generated in the history of our planet has probably been produced in the past several years by automated Big Data collection systems. Google’s search database alone collects literally billions of records on a daily basis and will presumably continue to do so into the foreseeable future, accumulating an almost impossibly large amount of organic information. Prewitt (2013: 229) refers to this as a “digital data tsunami.” Survey data, by contrast, is many orders of magnitude more modest in volume.

**Data Structures**

By data structures, we mean the ratio of observations to variables. Big Data commonly have higher ratios (i.e., vastly more observation points than variables), and surveys have much lower ratios (i.e., many more variables but for vastly fewer observations). Prewitt (2013) describes survey data as case-poor-and-variable-rich, and Big Data as case-rich-and-variable-poor.

**Velocity**

Data velocity is the speed with which data is accumulated. Big Data’s velocity, of course, means that it can be acquired very quickly. Not so with surveys, which require greater planning and effort, depending on mode. Well-done telephone surveys can take weeks to complete, and well-done face-to-face and mail surveys can require months of effort. Even online surveys require at least several days of effort to complete all “field” work. Where government and business decisions must be made quickly, Big Data may increasingly become the only viable option for instant analysis. Indeed, many complex organizations now employ real-time
“dashboards” that display up-to-the-minute sets of indicators of organizational functioning and activity to be used for this purpose, and one of the stated advantages of Google’s Flu Index (to be discussed below) and similar efforts has been the almost real-time speed with which the underlying data become available, vastly outperforming surveys, as well as most other forms of data collection. Big Data is collected so quickly, without much in the way of human intervention or maintenance, that its velocity is sometimes compared to that of water emitting from a fire hose. Survey research will continue to have difficulty competing in this arena.

Data Chronicity

Data chronicity refers to data’s time dimensions. The chronicity of Big Data is much more continuous (or longitudinal) than that of most common cross-sectional surveys. With few exceptions, survey data are almost invariably collected over relatively short time intervals, typically over a matter of days, weeks or months. Some data collection systems for Big Data, in contrast, now are now systematically collecting information on an ongoing, more or less, permanent basis. There is an often incorrect assumption that the methods, coverage and content of Big Data remains static or unchanging over time. In fact, Big Data systems are often quite changeable and hence there is a danger that time series measurements may not be reliable and/or valid.

Intended Purpose

Design-based survey data are collected to address specific research questions. There are few examples of Big Data being intentionally constructed for research purposes, mostly by governmental agencies interested in taking ongoing weather, environmental, etc., measurements. Most Big Data initiatives, however, seem driven by commercial interests. Typically, researchers have a good deal of control over the survey data they collect, whereas most analysts of Big Data are dependent on the cooperative spirit and benevolence of large corporate enterprises who collect and control the data that the researchers seek to analyze.

Relative Advantages of Big Data

The main advantages of Big Data over survey data collection systems are costs and data completeness.

Costs of Data Collection

As mentioned earlier, Big Data has an important advantage in terms of data collection costs. Surveys, particularly those using an interviewer-assisted mode, continue to become increasingly expensive, whereas the costs of using available Big Data collected for other purposes may be less expensive. The cost of original collection of Big Data, though, is often very high. As research funding becomes more difficult to obtain, the economic attractiveness of Big Data make it difficult to not seriously consider it as an alternative data source.

Data Completeness

Missing data at both the item and unit levels is a difficult problem in survey research and the errors associated with it preoccupy many researchers. Big Data sets do not typically share this problem. Because most Big Data sets are based on varied data collection systems that do not rely directly on the participation of volunteers, and subjects are typically not even aware that they are contributing information to Big Data systems (on this point, see the section on Ethical Oversight...
below), non-observations due to failure to contact individuals, or to their unwillingness or inability to answer certain questions, or to participate at all, is not a problem. But Big Data is also not perfect, as we would expect for example that monitors and other recording devices will occasionally malfunction, rendering data streams incomplete. As with surveys, the information missing from Big Data sets may also be biased in multiple ways.

**Relative Advantages of Survey Research**

Advantages of survey research data over Big Data include its emphasis on theory, the ease of analysis, error assessment, population coverage, ethical oversight and transparency.

*The Role of Theory*

Some have argued that the we are facing “the end of theory,” as the advent of Big Data will make “the scientific method obsolete” (Anderson, 2008). Although some of the survey research reported in the popular news media is descriptive only, much of the research conducted using survey methods is theory-driven. Survey data are routinely employed to test increasingly sophisticated and elaborate theories of the workings of our social world. Rather than allowing theory to direct their analyses, Big Data users tend to be repeating some earlier criticisms of empirical survey research by inductively searching for patterns in the data, behaviors that left earlier generations of survey researchers vulnerable to accusations of using early high-speed computers to go on “fishing expeditions.” Fung (2014) criticizes Big Data as being observational (without design) and lacking in the controls that design-based data typically collect and employ to rule-out competing hypotheses.

*Ease of Analysis*

The sheer size of many Big Data sets and their often unstructured nature make them much more difficult to analyze, compared to typical survey data files. There are numerous packaged data management and statistical analysis systems readily available to accommodate virtually any survey data set. Big Data, in contrast, typically requires large, difficult-to-access computer systems to process, and there is a shortage of experts with the knowledge and experience to manage and analyze Big Data (Ovide, 2013).

*Measurement Error*

The error sources associated with survey data are reasonably well understood and have been the subject of robust, ongoing research initiatives for many decades (Groves et al., 2009; Sudman and Bradburn, 1974). We know that the Literary Digest poll was taken down by several error sources, such as coverage and nonresponse errors, that are well understood (Lusinchi, 2012; Squire, 1988). Errors associated with Big Data, however, are not well understood and efforts to systematically investigate them are only now beginning. Prewitt (2013: 230) observes that “there is no generally accepted understanding of what constitutes errors when it is machines collecting data from other machines.” Measurement error is an important example. Survey measures are typically the subject of considerable research and refinement, with sophisticated methodologies readily available for the design, testing, and assessment of measurement instruments (Madans et al., 2011; Presser et al., 2004). Big Data shares many of the challenges of secondary analyses of survey data in which specific indicators of the construct(s) of interest may not always be available, challenging the analyst’s creativity and cleverness to sometimes “weave a silk purse from a sow’s ear.” Indeed, those analyzing Big Data must work with what is available to them...
and there is seldom an opportunity to allow theory to drive the design of Big Data collection systems. There is also concern that those who generate Big Data are sometimes unwilling to share details of how their data are collected, to provide definitions of the terms and measures being used, and to allow replication of measurements and/or analyses based on their measurements.

One interesting example is the Google Flu Index that was mentioned earlier. In 2009, a team from Google Inc. and the Centers for Disease Control and Prevention (CDC) published a paper in Nature that described the development of a methodology for examining billions of Google search queries in order to monitor influenza in the general population (Ginsberg et al., 2009). They described a non-theoretical procedure that involved identifying those Google search queries that were most strongly correlated with influenza data from the CDC; hundreds of millions of models were fit during the development of the flu index. They reported the ability to accurately estimate weekly influenza within each region of the U.S. and to do so with only a very short time lag. In 2009, the flu index underestimated a non-seasonal outbreak, and researchers speculated that changes in the public’s online search behaviors, possibly due to seasonality, might be responsible (Cook et al., 2011). Despite an ongoing effort to revise, update and improve the predictive power of Google Flu Trends, it also greatly overestimated influenza at the height of the flu season in 2011-12 (Lazer et al., 2014a) and especially in 2012-13 (Butler, 2013). Lazer et al. (2014a) also demonstrated that Google Flu Trends had essentially overestimated flu prevalence during 100 of 108 weeks (starting with August 2011). A preliminary analysis of the 2013-14 season suggests some improvement, although it is still overestimating flu prevalence (Lazer et al., 2014b).

Couper (2013) has made the interesting point that many users of social media, such as Facebook, are to some extent motivated by impression management, and we can thus not be certain of the extent to which information derived from these sources accurately represents the individuals who post information there. Social desirability bias would thus appear to be a threat to the quality of Big Data as well as survey data. The fact that a significant proportion of all Facebook accounts, for example, are believed to represent fictitious individuals is another cause for concern. One estimate from 2012 suggests the number of fake Facebook accounts may be as many as 83 million (Kelly, 2012). Hence, concerns with data falsification also extend to Big Data.

**Population Coverage**

The Literary Digest Poll was big, but many believe it did not provide adequate coverage of the population to which it was attempting to make inferences. Rather, it likely over-represented upper income households with political orientations decidedly unrepresentative of the Depression Era U.S. citizenry. Clearly, volume could not compensate for or fix systematic error. Big Data faces similar problems. For Big Data that captures online activities, it is important to be reminded that not everyone is linked to the internet, not everyone uses Google search engines, Twitter and Facebook. Among those who do interact with the web, the manners in which they do are very diverse. The elderly, who are less likely to engage the internet, are particularly vulnerable to influenza, yet none of the Google Flu Index papers referenced here address this.

---

1 In 2008, a team of academic investigators and Yahoo! Employees published a similar paper (Polgreen et al., 2008). That team, however, had not continued to report on this topic.
issue. A related concern is the problem of selection bias. As Couper (2013) has observed, Big Data tends to focus on society’s “haves” and less so on the “have-nots.” In addition, in Big Data there can be a problem with potential violations of the “one-person-one-vote” rule. As Smith (2012) has commented, a large preponderance of some social media activities, such as Twitter and Facebook, are the products of the activities of relatively small concentrations of individuals, further calling in to question the adequacy of their coverage. Indeed, many Big Data systems have what Tufekci (2014) refers to as a denominator problem “created by vague, unclear or unrepresentative sampling.” Others have expressed concerns regarding the danger that Big Data “can be easily gamed” (Marcus and Davis, 2014). Donald Campbell (1970) wrote more than 40 years ago about the corruptibility of social data as it becomes more relevant to resource allocation decisions. Marcus and Davis (2014) discuss several Big Data examples of this. Design-based, “small data” surveys, in comparison, go to great lengths to insure that their samples adequately cover the population of interest.

Ethical Oversight

Unlike survey researcher’s insistence on obtaining informed consent from respondents prior to data collection, and emphasis on the distribution of de-identified data only, many Big Data operations routinely collect identifying information without the consent, or even the knowledge, of those being monitored. In comparison to the careful ethical reviews and oversight academic and government-based survey research routinely undergoes, the ethical issues surrounding Big Data are not yet well understood or recognized. There is little transparency or oversight in Big Data research, much of it being conducted by private groups using proprietary data. Unfortunately, recent events, such as Facebook’s mood experiments (Albergotti & Dwoskin, 2014; Kramer et al, 2014; Verma, 2014), are reminiscent of some of the ethical transgressions of past generations that led to ethical review requirements for federally funded research. Some have called for a Big Data Code of Ethical Practices (Rayport, 2011). The National Science Foundation has recognized this need and launched a Council for Big Data, Ethics, and Society in early 2014 “to provide critical social and cultural perspectives on big data initiatives” (see: http://www.datasociety.net/initiatives/council-for-big-data-ethics-and-society/).

Transparency

Transparency of methods is central to the ability to replicate research findings. There is a need for greater and more general understanding of how Big Data sets are constructed (Mayer-Schönberger and Cukier, 2013). Big Data is not yet transparent, and most Big Data is proprietary and commercially controlled, and the methods employed to analyze these data are seldom described in a manner that would facilitate replication. In fact, commercial interests often dictate against transparency. The Google Flu Index, for example, has never revealed the 45 or so search terms it uses to make its prevalence estimates. Laser et al. (2014b) have accused Google of reporting misleading information regarding the search terms they employ. While survey research is far from perfect when it comes to transparency of methods, there is general recognition of its importance. Professional journals demand disclosure of survey methods. Several years ago, AAPOR launched a Transparency Initiative, designed “to promote methodological disclosure through a proactive, educational approach that assists survey organizations in developing simple and efficient means for routinely disclosing the research methods associated with their publicly-released studies” (see: http://www.aapor.org/Transparency_Initiative.htm#.U97jXPk7uSo). In addition, codebooks, methodological reports, and other forms of documentation are standard by-
products of any reputable survey, and have been so for many decades. The documentation requirements of social science data archives, such as the Inter-University Consortium of Social and Political Research (ICPSR; see http://www.icpsr.umich.edu/icpsrweb/content/deposit/guide/chapter3docs.html) are very stringent. Documentation of internet data, by comparison, is extremely limited (Smith, 2013).

**Supplement or Substitute?**

Lazer and colleagues (2014a: 1203) have coined the term “Big Data Hubris” to refer to “the often implicit assumption that big data are a substitute for, rather than a supplement to, traditional data collection and analysis.” Others share this sentiment. The British sociologists Savage and Burrows (2007: 890) have considered the historicity of survey research and suggest that its “glory years” were between 1950-1990. Taking the long view, one has to wonder as to whether or not surveys might merely represent one of the first generations of social research methods, destined to be replaced by more efficient methodologies in an increasingly digital world? Just as the horse-drawn carriage was replaced by more advanced forms of transportation, might we be now witnessing the passing of a traditional methodology?

Only time will tell. Big Data, in its current stage of evolution, though, does not appear capable of serving as a wholesale replacement or substitute for survey research. Even Savage and Burrows (2007: 890) acknowledge that there are some niches “in which the sample survey will continue to be a central research tool because of the limits of transactional data” (i.e., Big Data). They cite crime victimization surveys, which consistently demonstrate victimization rates well in excess of estimates derived from administrative records. There are no doubt many other examples. But, Big Data is an important new and highly valuable source of information about our social world, one with the potential to help us understand and address some problems, including many of those being addressed by other presentations at this conference. So how do we reconcile small surveys with Big Data?

Several observers see important opportunities for surveys and Big Data to supplements or adjuncts to one another (Butler, 2013; Couper, 2013; Marcus and Davis, 2014; Smith, 2011; 2012); for Big Data to contribute rich context to surveys, and for surveys to help make sense of patterns uncovered, but not well understood, in Big Data. Combining multiple data sources to take advantage of the strengths of each and to help compensate for the limits of each approach, seems to be what the future holds for these largely unique data resources. Smith and Kim (2014) have proposed a multi-level, multi-source (ML-MS) approach to reducing survey-related errors through a coordinated effort to more systematically link survey data with information from multiple auxiliary sources, including Big Data. These linkages would take place at each possible level of analysis, from high levels of geographies through unique paradata sources that are themselves by-products of survey data collection activities, such as contact attempts and even computer key-stroke data from interviewers and/or respondents (c.f., Kreuter, 2013). In addition to private Big Data, administrative data files from governmental sources would also be linked to develop better understandings of social phenomena and the strengths and limitations of the various data sources themselves. As the former U.S. Census Bureau Director Robert Groves (2011: 869) has commented: “combining data sources to produce new information not contained in any single sources is the future.”
References


When Big Data Are Not Complete: 
A Time Geographic Analysis of Taxicab GPS Traces

Brittany Wood*
Department of Geography
Florida State University
323 Bellamy Building
Tallahassee, FL 32306-2190
Bsw05@fsu.edu

Mark W. Horner
Department of Geography
Florida State University
323 Bellamy Building
Tallahassee, FL 32306-2190
mhorner@fsu.edu

*Corresponding Author

Abstract

Recent developments in technology surrounding Global Positioning Satellite (GPS) data have offered researchers a multitude of new ways to explore mobile objects in new and innovative ways. Travel behavior and movements are often analyzed from GPS data in order to assess patterns and routes traversed by different vehicles. Unfortunately, databases that store GPS data are often large, unstructured, and increasingly difficult to manage given the size and computing limitations involved in exploring such big data. Even though these mobility data can be large and approximate real time capture, they are often not enumerative and as a result, there are situations where we do not have complete position data for all possible locations.

As an approach to this problem, recent research has expanded upon the time geographic framework in developing a network-based time-geographic density estimation (TGDE) metric to estimate probabilistic potential path trees (Downs, 2010). These metrics identify possible locations a moving object may or may not have passed between known origin, stops, and destination points. This paper seeks to utilize TGDE methods in order to investigate GPS trace characteristics taken from a large database of taxicabs over a specified time horizon. Data are taken from the San Francisco, California area. We present results of our movement analysis with TGDE and then discuss possible applications that can be explored using these metrics.
1. Introduction

Developments in Global Positioning Satellite (GPS) data have provided researchers interested in a variety of fields the ability to explore the movements and trajectories of various mobile objects (Zheng, et al., 2009). These GPS points are usually collected in a sequence with time stamps provided for each point. As a result, one is able to examine the many ways in which mobile objects move about locations. In practice, there are many datasets available that contain captured GPS point locations of moving vehicles, such as taxicabs, transit vehicles, or personal vehicles. Travel behavior and movements are often analyzed from GPS point data in order to assess patterns, routes, and stops for a given vehicle. However, databases that store collected GPS points are often large and unstructured to the point that they become cumbersome through which to sift. Relatedly, new methods in the field of time geography have been developed to analyze movement data (Chen et al., 2011; Winter & Yin, 2010). Even though these mobility data can be large and approximate real time capture, they are often not enumerative and as a result, there are situations where we do not have complete position data for all possible locations.

As an approach to this problem, recent research has expanded upon the time geographic framework in developing a network-based time-geographic density estimation (TGDE) metric to estimate probabilistic potential path trees (Downs, 2010). These metrics identify possible locations a moving object may or may not have passed between known origin, stops, and destination points. This paper seeks to utilize TGDE methods in order to investigate GPS trace characteristics taken from a large database of taxicabs over a specified time horizon. Data are taken from the San Francisco, California area. The analysis explores taxicab GPS trace patterns by characterizing their movements with TGDE. Primarily we seek to identify locations and clusters within the city that are frequented the most along the taxis’ moving trajectories. We present results of our movement analysis with TGDE and then conclude with suggestions for future research.

2. Background and Concepts

Concepts rooted in classical time geography have proven to be very useful in exploring the modern-day location patterns of mobile objects. Hagerstrand (1975) suggested the idea of studying human activities taking into account both the time and spatial constraints of an individual’s movements. These recorded movements essentially offer a framework that can be
used to derive understanding about activity patterns across time and space (Hagerstrand, 1975). Following Hagerstrand’s classical time geography, Miller (2005), among others, notes the idea of the 3-d space-time prism where possible locations within a bounded continuous space could be estimated given space and time constraints for a set of two locations (Neutens, et al., 2007). These measures were originally deterministic in nature, defining the maximum extents a mobile object may have traveled given a time budget. When mapped in two dimensions, these extents allow visualization of the potential path area within a geo-ellipse providing an idea of all the spatial locations an item may have been (Miller, 2005). However, classical time geographic measures do not necessarily provide a way of estimating an object’s likelihood of being at a particular place. In other words, while they can measure the extent of time and space an object can traverse, they cannot approximate the relative likelihood of an object being at a certain location at a certain time given the space and time constraints. As such, with the advent and popularity of geocomputational tools, classical time-geographic concepts have been utilized heavily in geographic information systems (GIS) in order to assess the spatiotemporal characteristics of individuals. A number of researchers have used the time geography framework in order to examine individual accessibility to goods and services (Delafontaine, et. al., 2012; Kwan & Weber, 2003; Kwan, 1999; Miller, 2007; Neutens et. al., 2012).

A network-based technique was developed in order to identify the paths most likely taken between two consecutive points on a network when the intermediate points are not known (Downs & Horner, 2012). Transportation analysts have used these methods in a variety of ways to investigate vehicle movements and their potential paths (Horner, et al., 2012; Song & Miller, 2014).

In recent years many large datasets containing GPS traces have been made publicly available. Unfortunately, these databases are often large and unstructured and sometimes exceed the capability of spatial computing approaches (Castro, et al., 2012; Savage, et al., 2010). Many researchers have developed algorithms designed to discover patterns within these large datasets. For example, Castro, et al. (2012) proposed a method to construct a model of traffic density based on large scale taxi traces, while Savage, et al. (2010) developed an algorithm which sought to mine frequent trajectories in large scale GPS datasets. This paper would extend past work by taking into account both a vehicle’s location and time, utilizing TGDE in order to identify not
only the most likely paths traversed within a large GPS trajectory dataset, but also identify the relative likelihood of which locations will be visited the most based on cumulative trajectories.

Originally Time Geographic Density Estimation was developed blending kernel density estimation and time geography in order to delineate a mobile objects possible locations in continuous space when the locations are not originally known (Downs, 2010). For example, given two known location points of a single object and where the time it takes to get from one point to another is used as a time budget, TGDE can be used to calculate the relative likelihood that an object was present at any location within the given space time constraints. Down and Horner (2012) extended this work in order to cultivate a network-based TGDE metric which is used to estimate probabilistic potential path trees, or the most likely path an object is likely to take. Essentially, the network based TGDE shows the locations a moving object was more likely to have passed through between known stops on a network, such as a roadway. The equation for network based TGDE is depicted below and follows the notation of Downs and Horner (2012):

\[
f_t(x) = (N - 1)^{-1} \sum_{i=1}^{N-1} PPT^* \left( \frac{t_p(l,i,x) + t_p(x,j)}{t(l,j) - t_a(l,j)} \right) s_{ij}^{-1}
\]

(1)

Here \( f_t(x) \) represents the time-geographic density estimation for any GPS point location, \( x \), on a network. \( N \) is designated as the number of control points within a specified dataset, while \( (i, j) \) are the consecutive points denoted in the same dataset. \( PPT^* \) is the distance weighting function for the potential path tree for each set of control points. The time variables are represented as \( t(,) \), which is the time spent between two control points, and \( t_a(,) \) is the time spent for an immobile activity between two location points. Lastly, \( t_p(,) \) is the minimum travel time between two point locations on a network. This appointed minimum travel time is estimated based on the maximum speed a vehicle or object may travel along the shortest path \( p \). Essentially, the travel budget between two control points is computed by dividing the shortest path travel time between control point \( i \) and any location \( x \) on a network to control point \( j \) by the recorded travel time between control points \( i \) and \( j \). TGDE values are then only assigned to locations that fall within an object’s overall potential path tree. In this case, linear decay is used as the weighting function since it is assumed that the TGDE values will decrease over space as an object’s potential path locations \( (x) \) get further from the original control points.
Finally, $s(i, j)$ is incorporated into the equation as a dimension of $PPT^*$ for each pair of control points. This parameter is used as a way to account for complexity that may exist in the network. Basically, potential path trees for different pairs of control points may vary in the number of possible routes between $i$ and $j$. As such, because intensity values are summed for each pair of control points, computed intensities would be higher for $(i,j)$ pairs with multiple possible paths compared to $(i,j)$ pairs with fewer or a single possible path. Essentially, the dimension of an $(i,j)$ pair is defined as the total number of possible shortest paths or all of the possible paths that can be reached from $i$ to $x$ to $j$ within the given time budget between a set of control points. The TGDE network intensity values are divided by this number $s(i,j)$ in order to account for these differences in complexity across the network, so that they may be adjusted to properly reflect in network space.

Since its formulation the TGDE network equation has been utilized to explore vehicle movements in network space in order to discover patterns surrounding travel behavior (Horner & Downs, 2014; Horner et al., 2012). Horner et. al. (2012), attempt to use TGDE in order to reconstruct possible paths from unknown location points that exist in travel surveys. Essentially, TGDE was used as a tool to estimate unknown locations and stops that were missing from a sample of travel survey data(Horner et al., 2012). Horner and Downs (2014) propose a TGDE accessibility metric that fuses both individual and aggregate modeling perspectives. This concept was expanded in a study conducted by Horner and Wood (2014) where a TGDE accessibility framework was developed in order to identify peoples’ individual food environments using information on travel movements and the location of food stores within a given study area. Results indicated that place-based and individual differences in access to food stores could be quantified and compared in order to identify patterns of inequity across the transportation system(Horner & Wood, 2014). To date, TGDE metrics have not been used to explore uncertainty in large GPS datasets. This paper proposes the TGDE framework as a tool in order to organize and explore patterns that may exist in big data.

3. Methods and Data

Taxicab GPS traces recorded in 2008 in San Francisco will be used for this project. This data set contains the GPS coordinates for approximately 500 taxis collected over 30 days in the San Francisco Bay area (Piorkowski, et al., 2009). For each taxi, there are approximately 25,000
GPS coordinates with a time range of 10-60 seconds between each point (Figure 1). Each taxicab was outfitted with a GPS receiver, which sends a timestamp, identifier, and geo-coordinates to a central server. For privacy reasons, no direct access to identifiable cab numbers is provided in the data set, but the data does detail whether or not a passenger is using the cab or not for specified time periods (Piorkowski et al., 2009).

The data sets needed some preliminary processing before conducting any analysis. Times were recorded in Unicode time and needed to be converted to regular time, and then time intervals between each point in seconds needed to be calculated for each set of control points. Occasionally, the GPS receiver would capture time intervals when the taxi was parked. For these instances, when time intervals between two control points were greater than 200 seconds, the taxi was assumed to be parked and not included in the analysis. Ultimately, we will use TGDE to explore patterns and possible unknown locations where we can interpolate incomplete location information in these ‘big data.’

Road network data for the San Francisco area was acquired from the United States 2010 Census Tigerline files (US Census Burea, 2010). A node-based network was created in a TransCAD (Caliper, Inc.) GIS environment estimating travel times based on road classifications using Census Feature Class Codes (CFCC), which provide information on travel speeds. Unfortunately, quick and accurate methods to computationally sync GPS traces exactly to road networks do not exist. Since recorded GPS coordinates do not sync up exactly to the road network, each GPS coordinate was spatially connected and associated with an already existing node on the road network. As such, there may be some minor location error associated with computed potential paths.
After the GPS coordinates are connected to its nearest node on the road network a travel time matrix was computed in TransCAD recording the travel times to and from all nodes existing in the network. Using GISDK, Caliper’s development language, equation (1) was coded in order to automate the time geographic intensity values. Intensities computed by the TGDE formula are all node-based in that each intersection on the network represents a possible location ($x$) that a taxi may have traveled to in between its origin ($i$) and destination ($j$) points.

The script written to automate the TGDE results was applied to all the GPS points recorded consecutively for a given taxi. First a test was conducted on a small sample of points in order to determine whether the code could automate estimated intensity values accurately with the given input data. Figure 2 below depicts an example of the calculated TGDE values for one taxi over two hours. Black square boxes represent the actual locations of GPS points, while darker color shadings represent higher intensity values, or the locations that a taxi was most likely to have been along the network. Ultimately, we created a sequence of TGDE surfaces that...
depict intensity values for each taxi. In the next section, we will discuss how TGDE metrics can be an effective and useful way of visualizing and analyzing big data location patterns.

Figure 2. TGDE Intensity Values for One Taxi over Two Hours

4. Results and Summary

Using the spatial and temporal characteristics of each GPS coordinate collected over 30 days for approximately 500 taxis, we will be able to account for the uncertainty that exists in between GPS point locations for each taxi along the network. We observed that TGDE was adequate in determining the most likely locations a taxi would possibly frequent over an extended period of time. However, this research extends past work by examining multiple trips within a large GPS dataset and estimating the most likely locations a taxicab may have visited over time. Essentially, the TGDE technique allows us to visualize where the vehicle most likely was between known GPS point locations (Figure 3). Darker color shadings indicate higher intensity values, while lighter colors indicate lower intensity values. Intensity values are indicative of the likelihood of taxicab presence. As the figure depicts below, darker intensity values are mostly located in the upper right corner of the image, suggesting that this particular taxi spends most of its time in that particular geographic area.
Examining the time-geographic density surface for one taxi and the nodes assigned the highest intensities for that taxi could provide insights into the most popular locations one taxi could visit over any specified time period (an afternoon, a day, or a month). Figure 4 below represents a TGDE surface estimation for one taxi over one month including the nodes with the five highest intensity values. These particular nodes are located in the same vicinity and indicate that this area is a frequent location visited by this individual taxi. In effect, TGDE has allowed us to quickly quantify and identify information from a large dataset that would otherwise have been cumbersome and computationally difficult to extract information.
With these TGDE intensity values we can quantify and visualize uncertainty within large GPS datasets. The TGDE approach can be used as powerful tool to develop explorative analyses on big data. For example, Horner and Wood (2012) used TGDE as a way of quantifying and determining individual and place-based accessibility metrics associated with local food environments. As we have shown, these metrics can be applied to the taxi data and would provide a way of estimating accessibility to any number of activities by taxi in the San Francisco Bay area. This could help policy makers and those interested in the transportation system hone in on any issues or inequities that may exist across the city in terms of transport.

Going forward, census data could also be blended with TGDE surfaces in order to obtain information about the population surrounding the locations with the lowest and highest intensities. Taxi stops can be determined by points in time when the taxi becomes unoccupied versus occupied. This would in turn determine where a customer vacates or boards a taxi.
Exploring the demographics of where individuals tend to be dropped off and picked up could offer general information about taxi clientele which would support marketing analyses. Multiple time-geographic density surfaces could also potentially be combined in order to show the most likely locations not one, but multiple taxis traverse over time. Identifying the most probable paths a vehicle takes over time could prove to be very useful for advertisers seeking to place advertisements at the most frequented locations or for transportation planners to identify road paths that need frequent maintenance. These TGDE surfaces could also be combined in order to visualize traffic hotspots for multiple taxis as a way of depicting the locations taxis frequent the most along their trips. Previous work using TGDE has been focused on identifying the most likely path for one vehicle trip, instead of many vehicle trips over time. This problem is exacerbated when datasets are large and time intervals between vehicles are non-trivial. Datasets for previous analyses have been considerably smaller and easier to manage. This paper builds upon this research by showcasing TGDE as a potential tool for big data visualization and management. Ultimately, TGDE proves to be a useful application in the effort to understand, visualize, and explore large GPS datasets.

Acknowledgements
The second author would like to acknowledge the National Science Foundation (NSF) for its support of portions of this research (BCS-1062924). The contents of this abstract represent the views of the authors and do not represent those of the NSF.

References


US Census Burea. (2000). *TIGER/Line Shapefiles and TIGER/Line Files*.


Considering smartphones: Testing big data veracity via user privacy and trust in location-aware applications

Dr. Caitlin D. Cottrill (Corresponding Author), Lecturer
University of Aberdeen
School of Geosciences
Department of Geography & Environment
St Mary’s Building
Elphinstone Road
Aberdeen, Scotland, UK AB24 3UF
Phone: +44 (0)1224 273694
c.cottrill@abdn.ac.uk

Abstract:
The rapid adoption of smartphones and the location-sensing applications for which they provide a platform has provided scope for efficient and dispersed collection of location information that may be beneficial to the field of transportation. The quality and comprehensiveness of such data, however, depend in large part on the attitudes of users towards these applications, their trust in the application developers and data users, and their practices relative to these applications in the context of their personal use. In this paper, we initiate an exploration of some of these issues using a stated preference and self-reported behavior survey of US-based smartphone users in order to generate an understanding of underlying issues that may impact the quality and quantity of data collected through these technologies.

Keywords: Privacy, Attitudes, Location Information
INTRODUCTION

The so-called ‘privacy paradox’ (whereby persons indicate a high concern for privacy but take few or no proactive measures to protect it) has been studied by researchers in the context of social media (Barnes, 2006), (Dwyer, Hiltz, & Passerini, 2007), marketing and purchasing behaviors (Norberg, Horne, & Horne, 2007), (Xu, Luo, Carroll, & Rosson, 2011), and general theories of behavior (Kehr, Wentzel, & Mayer, 2013). It is an area gaining widespread attention as we move towards the era of ‘Big Data,’ in which we assume that consistent, comprehensive, and accurate data will be collected on numerous aspects of daily life, including transport, purchasing behaviors, and online activities, amongst others. In such a context, the willingness of people to share information, whether known or unknown, is a key component of developing accurate models and predictions for increasing urban efficiency. The idea of unknown data collection is, however, beginning to raise worrying eyebrows amongst the population, particularly given the recent spate of revelations regarding the collection, sharing, and use of personal private data from such companies as Snapchat (FTC, 2014), Google (Cellan-Jones, 2014), and Facebook (Van Grove, 2014). Perhaps no revelation, however, was more damaging for perceptions of data privacy than that related to the United States National Security Agency’s (NSA) PRISM program, which ‘…targeted the Internet communications and stored data of “non-US persons” outside the US and those communicating with them, and [revealed] the extent to which US companies cooperate with the government (Landau, 2013).’ Details of the PRISM program were leaked to The Washington Post and Guardian newspaper by then-Booz Allen Hamilton contractor Edward Snowden, and published publicly on 6 June 2013.

Since the PRISM revelations, the topic of privacy has become omnipresent in discussions of the digital realm and Big Data. Though surveys conducted since the revelations have been inconclusive in their findings regarding the concerns experienced by members of the public, governments, and corporations (Intralinks, 2014), Pew Research Center 2013), few studies have been conducted that assess how people’s perceptions and behaviors regarding privacy and trust have changed since the initial information was made public. In this study, we use as a basis for exploration a survey on attitudes towards privacy and trust regarding the use of location-sensing smartphone applications, completed on 3 June 2013 – immediately prior to the initial release of information on the PRISM program. Using this study as a basis for comparison, we conducted a follow-up survey nearly a year later, using the same instrument, but with an added set of questions regarding respondents’
knowledge of the PRISM program and perception and self-reported behavioral responses in light of that knowledge. In conducting this comparative study, we hope to add value to explorations of how the increased attention to privacy in the general media may have influenced the public’s attitudes towards the privacy of their personal information.

In this paper, we begin by introducing the general concepts of information and location privacy, and how they have been perceived in both academic and popular literature. We then address how these concepts have gained increasing attention in light of new technologies that begin to blur the boundaries of the creation and ownership of ‘personal’ and ‘private’ data, and discuss how the mediating influences of trust and context may begin to impact upon their public perception. Following these discussions, we present the findings of the two surveys conducted in 2013 and 2014 that use the case of location-enabled smartphone applications (or ‘apps’) to test public perceptions of privacy in a pre- and post-PRISM environment.

**CONTEXT**

According to the Pew Research Center’s Internet & American Life Project, as of January 2014, 90% of American adults owned a cell phone, with 58% of those being smartphones (Fox, 2014). A separate 2012 Pew report found that of those Americans with smartphones, roughly 75% use some sort of location-based service (Zickuhr, 2012). The growing ubiquity of smartphones and attendant location-enabled apps has had widespread ramifications in the transportation sector, with increasing use of GPS-enabled smartphones for the dissemination of travel surveys, combined with the massive quantities of location and travel behavior data currently collected by location-enabled apps, ushering in a new era for the collection and mining of detailed, extensive transportation data sets. In addition, location-enabled technologies have allowed targeted transportation information such as bus arrival times and safe bicycle routes to be provided to users via use of specialized apps, further expanding the quantities and scope of data available on both travelers and the travel environment (Viktorsson, 2013).

Despite these gains, however, questions related to privacy and trust in the context of app use remain, particularly given that one of the expected benefits of emerging technologies is the ability to collect, share and use detailed, accurate location and mobility data over time. If trust concerns cause participants to reduce their data sharing, this may limit the ability of
Cottrill planners, engineers, and others to realize the full potential benefits expected from the use of location-sensing and collecting technologies. Impacting issues, such as attitudes towards privacy and willingness to share and disseminate information, have the potential to greatly impact the quality of data collected via smartphone applications, and highlight the need to more fully evaluate the likelihood of data manipulation.

In this study, we examine emerging and changing trends of location-enabled smartphone app adoption and use, looking, in particular, at how attitudes towards potential concerns (i.e., privacy and trust) may impact, or be reflected in, patterns of behavior, particularly in light of the PRISM revelations. In light of increased media scrutiny of privacy and technology since the PRISM program was made public, we present the results of two surveys that examine in more detail how current users of smartphones make choices related to location-enabled app use and their attitudes regarding these app benefits and concerns. Findings will be used, first, to propose methods by which identified concerns may be adequately addressed by app developers and data users, and second, to evaluate how users may potentially respond to increased public attention to potential misuse of personal data.

We first review existing research on smartphone adoption rates and dissemination of location-based applications. Next, we examine the types of data that may commonly be gathered through use of these applications, and how they are treated within the public and private sectors regarding regulations on sharing and/or selling. Following this discussion, we report the findings of two surveys: the first a pilot survey of 120 U.S.-based smartphone users on their use of location-sensing apps, and their concerns and attitudes regarding trust and use of these technologies; and the second the same survey repeated roughly one year later. Results from these surveys will be examined and compared, and areas of needed future research identified to ensure that user concerns are addressed in relation to location-based smartphone applications such that the data gathered through use of these apps may itself be trusted. It is anticipated that this study will provide an initial analysis of areas of concern to be examined as we move towards the use of ubiquitous technologies for support of transport planning and projects.

BACKGROUND

While standard mobile phones were available in the marketplace by the early 1980s, the commercially viable smartphone is generally regarded as having emerged in the early 2000s. According to PC Magazine, a smartphone can be defined as, “A cellular telephone with built-in applications and Internet access. In addition to digital voice service, modern smartphones
provide text messaging, e-mail, Web browsing, still and video cameras, MP3 player and video playback and calling. In addition to their built-in functions, smartphones run myriad free and paid applications, turning the once single-minded cellphone into a mobile personal computer (PCMag.com, n.d.).” The rapid adoption rate of smartphones has relied in large part on the widespread implementation of the underlying infrastructure necessary for their effective operation, including global positioning systems (GPS), wireless network access points for Wi-Fi, and 3 and 4G (third and fourth generation) networks. In addition, the importance afforded to the smartphone has also been reflected in growing investments made in research and development of these resources in the areas of government, private corporations, and public and private institutions (Andrews & de Serres, 2012), (The Insight Research Corporation, 2012)

In addition to investments being made in technological and infrastructure improvements in support of smartphone deployment, many resources are also being invested in development for smartphones. The market for mobile applications (“apps”, or small programs designed to run on mobile computing platforms) has been growing rapidly, with recent estimates showing the Google Play/Android marketplace with roughly 1.3 million apps (AppBrain, 2014) and the Apple App Store with roughly 1.25 million (Parfitt, 2014) available for smartphones and tablet computers. The purchase and/or download of these apps has also been growing rapidly – as of September 2013, IT research company Gartner estimated that mobile app stores would see app downloads reaching 102 billion in 2013 (Gartner, 2013).

It is evident from these figures that the use of smartphones and attendant apps has become fairly commonplace over the last several years. It should be noted, however, that differences are still evident in the sociodemographics of persons who use smartphones, in the types of applications they use, and in how they treat their personal information. For example, according to findings from the Pew Research Center (2013), younger, more educated, urban males tend to be more likely to use smartphones, though growth has been seen across most demographic sectors as smartphones have become more common, and as more developers enter the app market.

Such rapid adoption and dissemination, however, initiates concerns over the data that such apps collect, and with whom it is shared. As part of the Wall Street Journal’s “What They Know” series, they evaluated a set of 101 popular applications (50 each for Android and iPhone, along with the WSJ’s app for each) to determine what information was being collected, and with whom it was shared. While they tested a range of data types (including username and password; contacts, age and gender; location; phone ID; and phone number),
the findings regarding location data sharing were particularly revealing. For the iPhone, the study found that 26 of the 50 apps shared location data with third parties, while an additional six shared this data with the app developer (Thurm & Kane, 2010). For the Android, 21 apps shared location data with third parties, and an additional four with the app developer (Thurm & Kane, 2010). Such findings have a number of implications for the transport and mobility sectors. First, it provides an indication that a large number of applications are collecting and sharing location data, which may signal the potential for public transportation agencies to partner with application developers to access detailed, useful data on a sample of the population. At the same time, however, it indicates the potential lack of regulation overseeing this market sector, which may reduce the willingness of public agencies to tap into this data source. Such implications suggest the need to better understand the landscape of the use of location-aware applications, and how user adoption of and behavior regarding them may impact the quality of data collected, as well as the manner in which partnerships are made.

In the following section, we discuss the findings of two small surveys conducted in June of 2013 and May of 2014. The surveys used essentially the same instrument and methodology, the only difference being the inclusion of a set of questions related to the participant’s knowledge of and response to the NSA PRISM program regarding use and management of application data in the 2014 survey.

SURVEY OVERVIEW AND FINDINGS

Overview
In conducting the research related above, it was noted that there is little extant information available on trust in the context of location-sensing smartphone applications, and how this may impact or influence a user’s behavior regarding these applications. Because the manner in which a consumer interacts with an application may impact on the value of data gathered through these applications, and due to outstanding questions regarding data quality in the context of provenance, we determined to conduct a short stated preference survey to assess attitudes towards privacy in the mobile environment, if and how these attitudes may influence a consumer’s use of specific types of location-sensing apps, and if users have been proactive in maintaining or adjusting privacy settings in consideration of their concerns. We generally hypothesize that users will indicate that they have privacy concerns related to their use of location-sensing applications, but that this will not necessarily be reflected in their use of these applications in a way that would impact the quality of the location data from a planning
point of view. An initial survey was developed and tested by six pilot participants. Pilot responses led to the elimination of two questions, and additional definitions and examples given for a number of terms used in the survey (such as geo-fencing) in order to ensure clarity and comprehension. Following the increased media scrutiny of privacy issues following revelations regarding spying in the NSA PRISM program, we then conducted a repeat survey to determine if initial findings had changed in the interim.

For both surveys, we recruited approximately 120 participants (124 in 2014) using the Amazon Mechanical Turk (MTurk) marketplace. Each respondent in the 2013 survey was paid US$0.50 to complete an 18-item instrument. In the 2014 survey, each participant was paid US$0.40 to complete a 21-item survey. The incentive was reduced in the second iteration due to a reduction in the amount of time estimated to complete the survey. Initially, it had been estimated that the survey would take roughly 20 minutes complete; however, results from the first survey indicated an approximate completion time of 8 minutes and 49 seconds, an average time which increased to only 9 minutes 8 seconds for the second survey.

Both surveys were hosted on the website SurveyMonkey, and all participants were asked to confirm that they reside in the United States, are 18 years of age or older, and own a smartphone. We acknowledge that the use of MTurk for selecting our sample will create a biased sample population; however, as this is intended to provide an initial assessment of the landscape regarding smartphones and location privacy, we felt that the trade-off between obtaining a geographically dispersed sample of persons who are not affiliated with the researcher was more appropriate than obtaining a general convenience sample. The use of Mechanical Turk has been supported by researchers such as (Paolacci, Chandler, & Ipeirotis, 2010), who report that, despite some differences with the overall US population, MTurk respondent results are comparable to other Internet samples or samples obtained from college campuses; (Buhrmester, Kwang, & Gosling, 2011), who found that MTurk samples tend to be more demographically representative than traditional convenience samples or college-based sample sets; and (Berinsky, Huber, & Lenz, 2012), who found that MTurk respondents are generally representative of the US population, though potentially younger and slightly more liberal ideologically. If we estimate a US population of adults over the age of 18 as roughly 231,195,000, and accept that approximately 56% own a smartphone (Duggan & Rainie, 2012), this leaves our estimated population of interest as 129,469,200. A sample size of 120 provides us with a confidence level of 95% with a margin of error of +/- 8.95. While subject to bias from demographic discrepancies, we feel that this sample is adequate for an
initial assessment of attitudes towards privacy in the context of location-sensing applications. Figures 1 and 2 below provide a map of the geographic locations of respondents.

**FIGURE 1 Map of 2013 survey responses**

![Map of 2013 survey responses](image)

**FIGURE 2 Map of 2014 survey responses**

![Map of 2014 survey responses](image)

Though heavily weighted towards the northern and eastern portions of the United States, patterns and centers of the US population are reasonably well reflected in the location of respondents in both instances.

It is not possible to calculate the survey response rate, as we cannot track the number of persons who were eligible to participate in the survey and who also viewed the assignment. In the 2013 survey, of the 128 persons who consented to the requirements and began the
survey, 120 persons (set as the upper limit) finished the instrument, for a completion rate of 93.75%. The 2014 survey had a response rate of 92.3%. Respondents were asked a number of questions regarding their use of location-aware applications along with a number of demographic questions. For the second survey, participants were additionally asked if they had heard of the NSA PRISM program and, if so, if it had had any impacts on their concerns for privacy or their use of smartphone apps.

Findings

General Results

120 persons completed the 2013 survey and 124 persons completed the 2014 survey. Demographic characteristics of respondents for each survey, along with the differences in the two years, are shown in Table 1. While the response population skews slightly younger, wealthier, and more highly educated than the general US population, it is roughly reflective of patterns of smartphone ownership in the United States, as reported in (1). In order to determine if the overall samples were comparable, a two-sample F-test was conducted on each demographic to test for variances; with the exception of gender (f = 10.5), all demographic characteristics demonstrated fairly equal underlying population variances. The more male slant of the 2014 survey might be expected to have an impact on reported privacy concerns, as some studies have shown males to have fewer privacy concerns and be less proactive regarding privacy protection than females (Hoy & Milne, 2010); (Mao & Zhang, 2014); however, results from other studies have been inconclusive (C. D. Cottrill & Thakuriah, 2010), thus we feel that the surveys overall are able to be compared. Participants reported using a variety of different smartphone types, with most reporting using Android (56% in 2013 and 59% in 2014) or iPhones (41% in 2013 and 36% in 2014).

<table>
<thead>
<tr>
<th>Demographic</th>
<th>% Responses 2014</th>
<th>% Responses 2013</th>
<th>% Difference 2013-2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18-20</td>
<td>7.1%</td>
<td>13.8%</td>
<td>6.7%</td>
</tr>
<tr>
<td>21-29</td>
<td>46.0%</td>
<td>46.3%</td>
<td>0.3%</td>
</tr>
<tr>
<td>30-39</td>
<td>24.6%</td>
<td>24.4%</td>
<td>-0.2%</td>
</tr>
<tr>
<td>40-49</td>
<td>16.7%</td>
<td>10.6%</td>
<td>-6.1%</td>
</tr>
<tr>
<td>50-65</td>
<td>5.6%</td>
<td>4.9%</td>
<td>-0.7%</td>
</tr>
</tbody>
</table>
Participants were next asked the following question:

- Do you use any of the following types of location-sensing apps or services on your smartphone?
  - Navigation (such as Google Maps, Mapquest, Waze)
  - Public transport information (such as CTA Bus Tracker, TfL Journey Planner, AllSubway)
  - Other transport information/services (such as GetTaxi, mytaxi, BikeMap)
Place annotations and discoveries (such as banjo, Now, c:geo (geocaching app), AroundMe)

Geo-located social networking (i.e. apps whose uses are enhanced by using your geographic location) (such as Instagram, Foursquare, Loopt, Twitter, Facebook)

Travel (such as TripAdvisor, Airbnb, Travelocity, Kayak, LonelyPlanet)

Recommendations (such as Yelp, Zagat, Yell)

Family tracking or “Geo-fencing” (i.e. defining a virtual perimeter of interest in the real world) (such as GPS Tracking, FamilyTracker, NearParent)

Health and Fitness (such as RunKeeper, Map My Run/Ride, Pedometer)

Mobile commerce (such as Shopkick, Groupon, Living Social)

Local weather (such as Weather Channel, AccuWeather, WeatherBug)

Generally, this question was designed to understand the overall use of location-aware applications in order to obtain a ground-level understanding of how acceptable or applicable location-sensing apps are to smartphone users. In the 2013 survey, it was found that survey respondents use an average of five of the location-aware application types surveyed (SD=1.9 applications), with no statistically significant difference found between usage by gender.

Results of this question, broken down by smartphone type, are shown in Table 2.

**TABLE 2 Location Application Use by Phone Type (2013)**

<table>
<thead>
<tr>
<th>Application Type</th>
<th>Phone Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation</td>
<td>Android</td>
</tr>
<tr>
<td>Transit Information</td>
<td>64</td>
</tr>
<tr>
<td>Other Transport Information</td>
<td>12</td>
</tr>
<tr>
<td>Place Annotation</td>
<td>11</td>
</tr>
<tr>
<td>Social Networking</td>
<td>45</td>
</tr>
<tr>
<td>Travel</td>
<td>20</td>
</tr>
<tr>
<td>Recommendations</td>
<td>25</td>
</tr>
<tr>
<td>Geo-Fencing</td>
<td>8</td>
</tr>
<tr>
<td>Health &amp; Fitness</td>
<td>28</td>
</tr>
<tr>
<td>Mobile Commerce</td>
<td>31</td>
</tr>
<tr>
<td>Weather</td>
<td>59</td>
</tr>
<tr>
<td>Total</td>
<td>116</td>
</tr>
<tr>
<td>% Total</td>
<td>97%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>iPhone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation</td>
<td>48</td>
</tr>
<tr>
<td>Transit Information</td>
<td>9</td>
</tr>
<tr>
<td>Other Transport Information</td>
<td>4</td>
</tr>
<tr>
<td>Place Annotation</td>
<td>9</td>
</tr>
<tr>
<td>Social Networking</td>
<td>35</td>
</tr>
<tr>
<td>Travel</td>
<td>21</td>
</tr>
<tr>
<td>Recommendations</td>
<td>33</td>
</tr>
<tr>
<td>Geo-Fencing</td>
<td>8</td>
</tr>
<tr>
<td>Health &amp; Fitness</td>
<td>28</td>
</tr>
<tr>
<td>Mobile Commerce</td>
<td>25</td>
</tr>
<tr>
<td>Weather</td>
<td>45</td>
</tr>
<tr>
<td>Total</td>
<td>104</td>
</tr>
<tr>
<td>% Total</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Windows</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation</td>
<td>1</td>
</tr>
<tr>
<td>Transit Information</td>
<td>0</td>
</tr>
<tr>
<td>Other Transport Information</td>
<td>0</td>
</tr>
<tr>
<td>Place Annotation</td>
<td>0</td>
</tr>
<tr>
<td>Social Networking</td>
<td>0</td>
</tr>
<tr>
<td>Travel</td>
<td>1</td>
</tr>
<tr>
<td>Recommendations</td>
<td>1</td>
</tr>
<tr>
<td>Geo-Fencing</td>
<td>0</td>
</tr>
<tr>
<td>Health &amp; Fitness</td>
<td>1</td>
</tr>
<tr>
<td>Mobile Commerce</td>
<td>1</td>
</tr>
<tr>
<td>Weather</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
</tr>
<tr>
<td>% Total</td>
<td>6%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>BlackBerry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Navigation</td>
<td>3</td>
</tr>
<tr>
<td>Transit Information</td>
<td>0</td>
</tr>
<tr>
<td>Other Transport Information</td>
<td>0</td>
</tr>
<tr>
<td>Place Annotation</td>
<td>0</td>
</tr>
<tr>
<td>Social Networking</td>
<td>3</td>
</tr>
<tr>
<td>Travel</td>
<td>0</td>
</tr>
<tr>
<td>Recommendations</td>
<td>1</td>
</tr>
<tr>
<td>Geo-Fencing</td>
<td>0</td>
</tr>
<tr>
<td>Health &amp; Fitness</td>
<td>2</td>
</tr>
<tr>
<td>Mobile Commerce</td>
<td>1</td>
</tr>
<tr>
<td>Weather</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
</tr>
<tr>
<td>% Total</td>
<td>10%</td>
</tr>
</tbody>
</table>

The 2014 survey showed an overall decline in the use of the types of applications surveyed, with participants reporting an average of 4.3 types of app used. Table 3 below presents the summarized 2014 findings by phone type.
From these tables, it is apparent that surveyed users are most inclined to use navigation, social networking and weather applications. In part, this may be due to the likelihood of having applications pre-installed on purchased smartphones. Navigation and weather apps are particularly likely to come built-in to current smartphones, perhaps accounting for the large number of affirmative responses to these categories. No distinction, however, was made between apps that had been selectively installed by the respondent and those that were included with the phone at the time of purchase. While absolute numbers of application use dropped for all categories excepting Travel and Place Annotation between 2013 and 2014, these drops were not statistically significant.

**Attitudes towards privacy and trust**

All respondents in both surveys were asked to respond to the question, “In general, how concerned are you about privacy?” Table 4 shows the responses obtained in 2013 and 2014, along with the amount of difference.

**Table 4: Degree of Privacy Concern Reported by Survey Respondents 2013 & 2014**

<table>
<thead>
<tr>
<th>Privacy Concern</th>
<th>2013</th>
<th>2014</th>
<th>Change 2013-2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not concerned at all</td>
<td>5</td>
<td>4%</td>
<td>2</td>
</tr>
<tr>
<td>Not very concerned</td>
<td>23</td>
<td>19%</td>
<td>5.6%</td>
</tr>
<tr>
<td>Somewhat concerned</td>
<td>50</td>
<td>42%</td>
<td>50</td>
</tr>
<tr>
<td>Very concerned</td>
<td>30</td>
<td>25%</td>
<td>33</td>
</tr>
<tr>
<td>Extremely concerned</td>
<td>12</td>
<td>10%</td>
<td>8</td>
</tr>
</tbody>
</table>

The distributions for both surveys are roughly in keeping with previous research (Louis Harris and Associates, 1981), (C. Cottrill & Thakuriah, 2011) indicating that the majority of persons generally fall into a category of privacy concern generally referred to as “privacy
pragmatists,” indicating that they are generally concerned about privacy, but are willing to trade privacy and personal information in return for benefits if these are perceived as reasonable. The general question about privacy concerns, in turn, was supplemented by questions pertaining to trust in the context of the specific location-aware smartphone app types that were surveyed.

Respondents were asked to respond to the following question about each surveyed app type, “Do you have any trust concerns about how the app developers/administrators or their marketers will use the information that you share with the apps you use?” Responses were categorized as No; Yes, a little bit; Yes, a fair amount; Yes, a lot; or Do not use. Figures 3 and 4 show reported degree of trust concerns with the application types surveyed here for 2013 and 2014:

**FIGURE 3 Degree of trust concerns noted by application users (2013)**
It should be noted that, though respondents were asked to report their concerns only for those types of applications that they have reported using, the number of persons who reported trust concerns for each type of application was higher than that reported in Tables 2 and 3 above. It is likely that persons at times also reported their trust concerns for those applications that they do not use. It is evident from both surveys that the greatest concerns come from social networking and mobile commerce sites. While general patterns of concern remained consistent from 2013 to 2014, means testing showed that some areas of trust (for example, in navigation apps) had changed significantly from 2013 to 2014.

**Actions reflecting privacy/trust awareness and concern**

In addition to evaluating survey participant’s attitudes towards trust of the location-aware applications that they use, we also wanted to determine if they participate in behavior (or believe themselves to have participated in behavior) that indicates a level of general concern for privacy and use of personal data. Because it is difficult to assess how participants have acted with regard to their personal information, we use questions regarding two types of behavior in order to gain a better understanding of potential actions taken that may reflect privacy and/or trust attitudes. The first action tested is that of reading privacy policies of installed and/or used applications. Such a question is an inexact measure of privacy concern, as participants may not recall if they read an initial privacy policy prior to commencing use of...
an app; additionally, as noted by (Thurm and Kane, 2010), many apps do not provide privacy policies for consumers to review. However, as noted by (14) privacy policies constitute the most consistent and common method by which consumers are notified of privacy practices by app providers, thus they are used as an indicator of privacy concern by participants.

For each location-aware app type used, respondents were asked if they had read the privacy policy prior to installation and use. Response options were:

- No
- Yes
- Did not have a policy
- Don’t remember
- Do not use

Responses, again shown as percentage of persons who responded other than “Do not use” are shown in Figures 5 and 6.

**FIGURE 5  2013 Participant responses to question on whether they read the privacy policy prior to installation and use of app**
FIGURE 6 2014 Participant responses to question on whether they read the privacy policy prior to installation and use of app

TABLE 5 Change in % of people reading privacy policies, 2013 – 2014

<table>
<thead>
<tr>
<th>App Type</th>
<th>2013-2014 Change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
</tr>
<tr>
<td>Navigation</td>
<td>-6%</td>
</tr>
<tr>
<td>Public Transit</td>
<td>-11%</td>
</tr>
<tr>
<td>Other Transport</td>
<td>-17%</td>
</tr>
<tr>
<td>Place Annotation</td>
<td>-15%</td>
</tr>
<tr>
<td>Social Networking</td>
<td>-9%</td>
</tr>
<tr>
<td>Travel</td>
<td>-11%</td>
</tr>
<tr>
<td>Recommendations</td>
<td>-12%</td>
</tr>
<tr>
<td>Geo-Fencing</td>
<td>-12%</td>
</tr>
<tr>
<td>Health &amp; Fitness</td>
<td>-8%</td>
</tr>
<tr>
<td>Mobile Commerce</td>
<td>-15%</td>
</tr>
<tr>
<td>Weather</td>
<td>-2%</td>
</tr>
</tbody>
</table>

Here, it is notable that the vast majority of persons who responded to this question for each app type indicated that they did not read the applicable privacy policy, though the general trend from 2013-2014 indicates an overall decline in persons reporting that they had not read a privacy policy, this decline was largely accounted for by the number of people reporting that they “Don’t remember” if they read a privacy policy (as shown in Table 5). In both years, it is evident that persons tended to be more cautious with regard to social networking apps, with roughly 23% and 24% of respondents indicating that they had read the privacy policy. Also of note, however, are the number of persons who indicate either that the app type did not have a privacy policy, or that the respondent does not remember if the app had a
privacy policy (again, the latter of which showed an overall increase from 2013 to 2014). There was, however, an increase in the number of persons who report having read the privacy policy for Mobile Commerce apps, a category that was particularly low in 2013. As seen above, survey respondents in both years indicated less trust respective to these types of applications (66% reported some degree of concern in 2013, and 61% in 2014). In 2013, 3% of respondents indicated a lack of privacy policy from these types of application, while 13% indicated that they did not remember if these apps had a privacy policy. In 2014, the overall uptick in persons reporting having read these policies indicates that there may be an increased concern in how these apps treat private information. However, there are still indications that privacy policies are not adequate methods by which application providers may engender trust in their services.

Next, participants were asked to report on proactive measures they have taken regarding their trust concerns. The question was asked as follows:

- If you answered that you have some trust issues with one or more of the location-enabled apps that you use, does this impact any of your actions regarding these apps?
  - My concerns do not impact how I use these apps
  - I use a fake name for these apps
  - I do not use these apps to post pictures of myself, my friends or family
  - I do not use these apps to tag pictures of myself, my friends or family
  - I do not provide my home location information
  - I do not provide my current location information (i.e. where I am now)
  - I limit my sharing only to those in my social network
  - I do not provide my financial information to purchase goods or services through these apps
  - I use a separate email account set up specifically for these apps
  - I use these apps to receive information, but do not provide any information (such as reviews or recommendations)

Here, we are interested in determining if the self-reported concerns that users have about the apps that they use impacts the quality and/or type of data likely to be collected from their use. This is of particular interest in the area of transport as (as reported above) many agencies and organizations are exploring the use of either specifically-designed applications, or of partnerships with other collecting organizations, to supplement or enhance their ongoing data collection procedures. By examining some of the actions that people take with respect to
these applications, it is hoped that we can identify if any concerns or awareness should be raised relative to potential issues with data quality based on participant use.

Results of this question for both the 2013 and 2014 surveys are shown in Figure 7 below. It should be noted that of the persons who reported that their trust concerns have no impact on how they use specified applications, several also reported taking additional actions to limit data sharing. It is presumed that this indicates that participants have taken different approaches to different application types.

**FIGURE 7 Impacts of trust concerns on application use**

It is evident that most respondents take no proactive actions to protect their privacy within apps that they use, though many users reported both no impact and one or more specific actions. Responses indicated that survey participants were more likely to limit the amount of information and with whom they shared this information rather than providing fake or limited background information, though reported use of all actions excepting limiting the sharing of home location has dropped since 2013. This may indicate that persons using location information collected or approximated from location-sensing applications may generally be able to trust the provided information, but should be cautious about assuming that collected information is complete.

As a follow-up, survey participants were also asked if they have made any of the following privacy changes to any of the location-sensing applications that they use:

- I have turned off GPS access for some or all apps
• I have specified with whom my location and data can be shared for some or all apps
• I have turned off geotagging (i.e. automatic detection of location, which is added to the photo) on my phone's camera
• I have limited those apps that have access to my contacts, calendar or reminders
• I have limited those apps that have access to photos I have stored on my phone
• I do not share my mobile phone number with these apps

In contrast to the above, these questions are intended to determine if the user has changed or amended any settings within the app itself. Results are shown in Figures 8 and 9 below.

**FIGURE 8 2013 Application data access limitations**

**FIGURE 9 2014 Application data access limitations**
Here, it should be noted that settings for location-specific data services (i.e. GPS and geotagging) have the lowest likelihood of being adjusted, likely in relation to the need for these to be available to obtain the full benefits of these services. However, as above, sharing of these data with others (particularly those not in the respondent’s social network) is likely to be limited. Of note is that changes are inconsistent from 2013 to 2014.

Finally, participants were asked about their degree of trust (on a 4-point range from “Do not trust” to “Trust completely”) of app developers and administrators with respect to the following:

- Not to sell my data to outside companies without my explicit consent
- Not to provide my data to government or law enforcement officials
- Not to track my specific location without my explicit consent
- Not to access my contact list without my explicit consent
- Not to store my data for an unreasonably long amount of time
- Not to maintain, sell or share inaccurate information

Results of this question are shown in Table 6 below.

It is evident here that survey participants generally display little to no trust of app developers and administrators regarding the access and/or use of their data, and that this trust has generally declined from 2013 to 2014. Of note, as well, is that most participants indicated a similar or exact amount of trust regarding each indicator studied. These findings indicate a general level of distrust that remains relatively stable across all identified components. Of note is that there is generally less trust shown towards the actions of selling data, tracking the respondent, or providing information to government or law enforcement agencies. Respondents indicated a slightly higher degree of trust that a developers and/or administrations will refrain from accessing contacts or other data stored within the phone.

**Impacts of the PRISM Program Revelations**

In the 2014 survey, a separate set of questions was asked to determine if participants were aware of the NSA PRISM program and, if so, if this awareness had impacted their privacy concerns and/or actions. The question was asked at the end of the survey, so as not to taint initial responses, or prompt additional awareness. Of all survey participants, roughly half
(52%) reported awareness of PRISM, and these respondents reported the following regarding their overall degree of privacy concern:

- I am less concerned about privacy now: 1.5%
- No change: 29.2%
- I am more concerned about privacy now: 69.2%

In addition, roughly 40% of those participants who reported awareness responded that PRISM had influenced them to make changes regarding how they use their smartphone apps with respect to privacy. 25 participants reported specific actions they had taken, including comments indicating the following:

- Reduced app use: 8
- Pay more attention to privacy policies and/or practices: 3
- Restriction of data access: 2
- Restriction of data provision/sharing: 4
- Overall increased awareness: 6
- Other data protection measures: 5

Specific comments included:

- “Since becoming aware of the scope of the NSA’s access, and privacy concerns in general, I now read what access I am giving apps very carefully, as well as reading privacy policies on new apps. I just pay a lot more attention now to what apps are accessing and what I am agreeing to.”
- “I’m just always mindful about the privacy of my content or the permission I give apps. I always keep in mind that my data and information is being freely given out to others if it’s on their servers.”
- “In general, I am more cautious about releasing information. The actions discussed in the study are just some of the changes, but I am more judicious with information in other contexts as well.”
- “I’ve stopped putting my location information into social networking apps.”

While the comments provided indicate that revelations regarding the PRISM program have influenced degrees of privacy concern, regression analysis revealed no significant correlation between awareness of the PRISM program and overall degree of privacy concern. Additionally, no significant differences were seen in reported trust attitudes or privacy-preserving actions between those reporting knowledge of PRISM and those who did not in the 2014 survey.
TABLE 6  Degree of trust in privacy-impacting actions by app developers and administrators

<table>
<thead>
<tr>
<th>Answer Options</th>
<th>2013</th>
<th>2014</th>
<th>Change 2013-2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do not trust</td>
<td>Trust a little bit</td>
<td>Trust a fair amount</td>
<td>Trust completely</td>
</tr>
<tr>
<td>Not to sell my data to outside companies without my explicit consent</td>
<td>39%</td>
<td>44%</td>
<td>15%</td>
</tr>
<tr>
<td>Not to provide my data to government or law enforcement officials</td>
<td>51%</td>
<td>30%</td>
<td>16%</td>
</tr>
<tr>
<td>Not to track my specific location without my explicit consent</td>
<td>50%</td>
<td>28%</td>
<td>16%</td>
</tr>
<tr>
<td>Not to access my contact list without my explicit consent</td>
<td>33%</td>
<td>37%</td>
<td>24%</td>
</tr>
<tr>
<td>Not to store my data for an unreasonably long amount of time</td>
<td>46%</td>
<td>31%</td>
<td>19%</td>
</tr>
<tr>
<td>Not to maintain, sell or share inaccurate information</td>
<td>42%</td>
<td>32%</td>
<td>22%</td>
</tr>
</tbody>
</table>
Discussion
Given both the rapid dissemination and adoption of smartphones and the wide array of location-sensing applications available, it is reasonable to anticipate that these tools have the potential for wide ranging benefits in the transportation sector. As shown above, however, some care and consideration needs to be taken as to how we treat data collected through these means. While the findings above indicate that current practices regarding the use of location-sensing (and, thus, location collection) applications allow for relatively consistent and comprehensive data to be collected, user trust concerns may begin to play more of a role in how users respond to data collection applications. The recent revelations regarding the National Security Administration’s PRISM program has the potential to both heighten awareness of data and location surveillance as well as erode trust in the companies that collect that data. The results indicated above, however, indicate that the privacy paradox still holds true – though survey participants reporting awareness of the PRISM program indicated heightened awareness and concern for privacy in some areas, they showed no significant differences in attitudes and practices separate from the overall survey sample. However, according to The Washington Post, “The NSA prides itself on stealing secrets and breaking codes, and it is accustomed to corporate partnerships that help it divert data traffic or sidestep barriers. But there has never been a Google or Facebook before, and it is unlikely that there are richer troves of valuable intelligence than the ones in Silicon Valley (U.S. Census Bureau, n.d.).” Such a contention is no less true about location-specific information, and data that is invaluable for transport planning, modeling, and programming purposes is no less valuable for purposes of tracking and marketing.

The indications of lack of trust in data collection, storage and use practices is revealing, as this may indicate that further actions will be taken regarding the protection of personal data as users become more confident in their ability to change settings and selectively set with whom data are shared. While this is beneficial from the point of view of privacy, it indicates the potential for smartphone-gathered location data to be evaluated carefully for indications of bias before it is used in modeling or planning applications. Correlations with the privacy-seeking behavior of users and their self-stated privacy preferences indicate that there is a class of travelers on whom it will remain more difficult to obtain comprehensive sets of travel behavior data, as such users may be disinclined to allow their data to be shared without reservations or amendment.

CONCLUSION
As shown above, the potential for collection and use of vast quantities of detailed location data from smartphones is great. The widespread adoption and use of this technology has the potential to allow for more accurate and efficient transportation planning and programming activities to occur. However, the successful use of such data is contingent upon the recognition that user attitudes and behaviors towards the collection, dissemination and use of such data must be taken into account as we strive to develop methodologies that reflect and address the concerns (particularly those regarding trust and privacy) identified. By developing transparent and open models of data access, sharing, storage and use, some of these issues may be overcome, allowing for more value to be added to transportation processes. It is hoped that we have here identified a number of questions for further analysis and study, particularly regarding data quality and applications as they apply to data collection from personal mobile devices. Inconsistencies in results, particularly in responses from persons with respect to impacts of the PRISM program, indicate that this is an area ripe for further exploration and analysis.
REFERENCES


True lies in big data: Detecting location spoofing based on geographic knowledge

Bo Zhao* and Daniel Z. Sui
Department of Geography
The Ohio State University
Columbus, OH, 43210

* Contact Information: zhao.662@osu.edu
True lies in big data: Detecting location spoofing based on geographic knowledge

Abstract: Location spoofing has been an issue of concerns in urban big data with multiple implications. The proliferation of location spoofing in recent years has stirred up the debate on the possible invasion of location-based deceptive advertisements, spams and cyber espionage on the one hand, and the use of location spoofing as an effective countermeasure to protect individual geo-privacy and national security on the other. However, these polarized views are not very helpful for us to understand the complexities of location spoofing, and as of today we still lack a robust method to detect location spoofing and a holistic framework for understanding its multifaceted implications. The primary goal of this paper is to develop a quantitative approach for detecting location spoofing and to contribute our understanding of this complex phenomenon.

We tested our approach for identifying counterfeit locations using millions of geo-tagged tweets, which is one of the major sources of big data related to human activities. The preliminary results indicate that the proposed approach can successfully detect certain types of location spoofing. Based on the empirical results, this study further examines the possible motivations of location spoofing, its uncertain nature, and its potential social implications. Rather than simply neglecting the spoofing phenomenon, this paper calls for directly confronting this thorny issue, especially when any arguments or policies are drawn from these location based big data. Only then can we promote more effective and trustworthy uses of big data. Keywords: location spoofing, big data, social media, Twitter, uncertainty

1 Introduction

The penetration of big data, especially when embedded with location-based services, has deeply affected the contemporary urban studies (Batty et al, 2012). It is quite common these days that location based big data can reference online personal identities or behaviors to their physical locations. At the same time, the location based big data has also brought about new challenges, ranging from protecting geo-privacy at the individual level (Friedland and Sommer, 2010; Vicente et al, 2011) to sustain urban and cyberspace security at the collective level (Valli and Hannay, 2010). Among all those potential challenges, location spoofing, an action to masquerade
as being elsewhere to confuse others, has been a long-standing but often ignored geographic practice (Warner and Johnston 2003; Lenders et al., 2008; Lee and Buehrer, 2010). Actually, the movie industry has a long history to conduct location spoofings or “Geopiracy” (a terminology for describing this false attribution of location in visual arts) as well. Movie buffs will recall one of the classics “Gone with the Wind” (1939) was shot in Hollywood, California, but the story actually took place in the antebellum South, and another example is the recent movie “Inception” (2010), which builds an urban landscape that can be folded, bended, and stretched. Nonetheless, this scene by no means exists in the real world (Secor, 2013). Indeed, to create an utterly different scene from the present geographical environment on one hand facilitates the filmmaker to reduce cost, realize impossible plots, and avoid endangerment, and the unjustifiably “geopiracy” inflicts harm to the film tourism as well as triggering locational prejudice on the other (Vogel and Gomides, 2008). Additionally, location spoofing is widely adopted in cyber-warfare for hiding whereabouts: the Chinese Cyber command (PLA Unit 61398) has meticulously spoofed their actual location whenever conducting cyber-espionage (Mandiant, 2013), while the NSA (National Security Agency) and its UK counterpart GCHQ (Government Communications Headquarters) have devoted considerable efforts to unveil the whereabouts of hackers/cyber-attackers who deliberately anonymize their IP, MAC or DNS addresses (Robinson, 2014). Facing these immediate challenges, it is very important to develop effective methods to detect location spoofing in big data and to evaluate their broader social implications.

Moreover, although efforts have been made to explore geographical dimensions of big data (Crampton et al., 2013; Kitchin, 2013), different kinds of spoofings related to information security (Jagatic et al., 2007; Jin et al., 2011), and the data quality assessments as well as the uncertainty of geographic data (Goodchild, 1993; Zhang and Goodchild 2002), this critical issue has received little attention by the existing GIScience literature, largely due to most existing studies about location based big data have taken geo-tags or latitude/longitude coordinates as the actual locations for granted (Java et al., 2007; Cheng et al., 2010; Stefanidis et al., 2013) rather than critically examining either the credibility of given locational information or the multiplicity
of ways that location/space is implicated when creating that geographic data (Crampton et al., 2013). Our knowledge about location spoofing is still fragmented and very limited. To our best knowledge, this study is the first systematic attempt to develop a quantitative method for detecting location spoofing phenomena in big data.

Inspired by Goodchild’s (2013) argument that geographic knowledge is essential for assessing the quality of big (geo) data, we plan to develop an approach for detecting location spoofings using a series of geographic facts. Among these adopted facts, a fundamental one is that any individual has to travel within the speed limit of available transportation vehicles. Since social media feeds, such as Twitter, Foursquare, and Facebook, are major sources of big data related to human activities, we have purposely designed an empirical study with millions of geo-tagged tweets to test the utility and effectiveness of the proposed approach. After singling out tweets with counterfeit locations, we also critically examine the potential causes and social implications of location spoofing.

The remainder of our paper is organized as follows: related work on location spoofing and geographic data quality assessment is elaborated in the second section. After a brief description in data harvesting, the concept of a baseline approach is presented in section three. Section four introduces the details of our proposed methodology of location spoofing detection, followed by results and discussions in section five. The last section contains summary and concluding remarks.

2 Location spoofing and geographic data quality assessment

Based on the Open Systems Interconnection (OSI) model of the Internet, location spoofing can take place within any or a combination of the seven OSI layers (which includes application, presentation, session, transport, network, data link and physical layer), while the spoofed information diffusing through these seven layers can be somehow related to location, such as the raw coordinates acquired by GPS receivers, the IP values when geo-referencing to their corresponding countries, cities, or even streets, the top-level domain (e.g., us, uk, cn, jp, kr, etc.) of an email address, the regional or city code of a telephone number and the textual place name
in a user’s profile or other web content. Hence, in a broader sense, location spoofing can be considered as a set of spoofing techniques to falsify different forms of locational information. Particularly, as far as location based big data, especially geo-enabled social media feeds concerns, the counterfeit locational information are mainly in three forms, in terms of (a) the textual and sematic geographic information in the content itself, for example, in the sentence “the building is near the campus”, “campus” is a piece of textual geographic information to describe a location, whereas “near” has a sematic meaning to represent the spatial relationship with the given location “campus”; (b) the structured place name listed in a user’s profile or other web forms; and (c) the geo-tag in the form of longitude/latitude coordinates.

Social media feeds as one of the major sources of big data related to human activities, similar to volunteered geographic information (VGI), are generated by vast amount of independent users/volunteers. To assess the quality of VGI, Goodchild and Li (2012) have put forward three approaches: in terms of crowd-sourcing, social, and geographic approaches. Firstly, the crowd-sourcing approach, relying on the Linus’s law -- “given enough eyes, all bugs are shallow” (Raymond, 2001) -- expects that a crowd can collectively converge on the truth; secondly, the social approach controls the data quality by a social hierarchy of trusted individuals; thirdly, the geographic approach inquiries the question how to possibly assert some purported locational information is geographically trustworthy. As Goodchild (2013, 283) stated in a follow-up paper, the answer lies in “the knowledge of the principles by which the geographic world is constructed, in other words, the principles embedded in geographic knowledge.” In this sense, a series of targeted geographic facts or principles, putting forward to distinguish the geographical impossibilities and inconsistencies, include but not limited to, as listed by Goodchild and Li (2012), Tobler’s First Law of Geography (Tobler, 1970; Sui, 2004), Horton’s Law (1945) on bifurcation ratio and Central Place Theory (Christaller, 1966). Different from the first two approaches, the geographic approach is the only method that can potentially be automated to assess and identify the suspected location spoofing without human intervention.

Similar to geographic data assessment, the approach to detect location spoofing also rests upon
the mechanism of these geographic inconsistencies. Inspired by the idea of a geographic approach, we argue that the geo-enabled social media feeds, as a primary source of VGI data, should definitely follow the basic geographic laws and principles. Thus, in the following sections, we propose an identification approach based on a series of geographic facts. Of course, a crowd-sourcing or social approach can also be developed to detect location spoofing, but they are beyond the scope of this paper.

3 Data harvest and a baseline approach for detecting location spoofing

Before elaborating the method in depth, we briefly describe the procedures we used for data harvest, our research objectives, and preliminary exploration of the data through a baseline approach.

3.1 Data harvest

We use Twitter feeds to facilitate us to bring about the research question and further to conduct the empirical study. One major reason is that Twitter is a prevailing social media app that has nearly globally distributed users, easily operated APIs, and plenty of geo-tagged “tweets” (a tweet refers to a short text message with no more than 140 characters). Thus, to get a representative sample of Twitter users, we harvest through Twitter’s public stream API for geo-tagged tweets, and a bounding box parameter within the URL request allows the responding tweets having geo-tags. For example, if adding a list value [-180, -90, 180, 90] to a request POST, the Twitter server will then keep sending back real-time tweets from users all over the world. This crawling method can be considered as random sampling from all active Twitter users. Therefore, using an open-source Python library Tweepy\(^1\) to access Twitter stream API within a two-day period (from February 2 to February 4, 2014), we have harvested around 3 million geo-tagged tweets, and simultaneously, these tweets can be stored into a spatial-enabled database (e.g., SpatiaLite, PostgreSQL with PostGIS, Oracle with Spatial extension, etc.).

---

\(^1\) Refer to [https://github.com/tweepy/tweepy](https://github.com/tweepy/tweepy)
Each tweet record in the database includes data fields about the content, metadata (id, retweet count and etc.), user’s profile (username, id, screen name and etc.), adopted language, source (the name of the client for tweeting), created-at timestamp, latitude and longitude. Having the precise coordinates, each tweet can be plotted on to the geographic surface (See Figure 1).

3.2 Baseline identification and initial results

As suggested in Section 2, we propose a well-accepted geographic fact to facilitate the identification procedure. That is, any individual, regardless of the selected travel mode (e.g., by foot, bicycle, car, train or airplane, etc.), has to travel within the speed limit of transportation vehicles. So, during the time interval of posting two pieces of geo-tagged information entities (e.g., posts, tweets, or statuses, etc.), if the mean speed is beyond the limit (e.g., the maximum speed limit of the fastest civilian airplanes), we can confidently infer that at least one geo-tag of the pair has been spoofed.

Since each pair in the raw data set, for now, has only one tweet, we have to find the other half. To make sure the other half is the latest geo-tagged tweet, we implement a depth-first search for all the tweets in the raw dataset. As a result, we have around 2.8 million (2,864,605) pairs of geo-tagged tweets. As described, the crawling strategy allows to extract a direct sub-graph of the
entire raw data with no pre-assumption to exclude any specific group of tweets.

So, we can formalize the research question as: given a set of pairs of information $S$ that are posted by a number $N$ of Twitter users $U$,

$$S = \{t_n(u) | 0 < n < N\}$$

A tweet pair is defined as,

$$t(u) = < l_1(u), l_2(u), \text{dist}(u), \text{dur}(u), v(u), \text{meta}(u)>, \text{in which,}$$

- $l(u) = (\text{latitude}, \text{longitude}, \text{created-at timestamp})$ – a point in space-time
- $\text{dist}(u) = \text{GreatCircleDistance}(l_1(u), l_2(u))$ – the great circle distance between two locations
- $\text{dur}(u) = \text{TimeInterval}(l_1(u), l_2(u))$ – time duration between sending the two tweets.
- $v(u) = \text{dist}(u) / \text{dur}(u)$ – average speed
- $\text{meta}(u) = (\text{e.g., username, source, language, etc.})$ – metadata

Given $V_{\text{max}}$ as the maximum velocity an average individual is capable of travelling. Then, $S_{ls}$ denotes the subset of $S$ that containing counterfeit locational information,

$$S_{ls} = \{t_n(u) | 0 < n < N \text{ and } v_n(u) > V_{\text{max}}\}$$

After scrutinizing the speed of various types of civil commercial planes currently in service such as Boeing, Airbus, McDonnell, we found the maximum cruise speed is around 1,000 km/h. In other words, an average individual cannot travel faster than 1,000 km/h. And consequently, the pairs of geo-tagged tweets can be identified through comparing the average velocity with the maximum velocity.

Based on the baseline identification, there are 11,427 pairs of geo-tagged tweets (around 0.4 percent of the entire sample) beyond the maximum speed limit. Among the 0.4 percent, a large proportion of these tweets are related to personal activities. The user who post tweets of this sort is considered as an individual based account (see Table 1).
Table 1 Different types of identified tweets

<table>
<thead>
<tr>
<th>Categories</th>
<th># Sources</th>
<th># Tweets</th>
<th>Sources</th>
<th>Usages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advertising Companies</td>
<td>112</td>
<td>5501</td>
<td>Dlvr.it, GitHubJobs, Yakaz, 4Job.co, Ivory Standard</td>
<td>Location based advertising about jobs, products, and services.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tenki, Keto Sea, SemanticEarth, everyEarthquake, QuakeSOS</td>
<td></td>
</tr>
<tr>
<td>NGOs</td>
<td>60</td>
<td>1957</td>
<td>Tenki, Keto Sea, SemanticEarth, everyEarthquake, QuakeSOS</td>
<td>Disaster (Earthquake, wave tides, hurricane)</td>
</tr>
<tr>
<td>Government/military</td>
<td>27</td>
<td>1270</td>
<td>City of Portland 911 feed, Baltimore 311</td>
<td>Traffic accidents/jams warnings, emergency notifications, Recruitment.</td>
</tr>
<tr>
<td>Agencies</td>
<td></td>
<td></td>
<td>Google, MarsBots, Radio Wave API</td>
<td>Testing, debugging, monitoring and etc.</td>
</tr>
<tr>
<td>Robots</td>
<td>22</td>
<td>546</td>
<td>Twitter for iPhone, Foursquare, Instagram</td>
<td>Diverse individual motivations</td>
</tr>
<tr>
<td>Individual based accounts</td>
<td>26</td>
<td>2153</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Compiled by authors

However, it is problematic and even a little risky to assume all of these 11,427 pairs of tweets as location spoofings, especially for those from non-individual based accounts. Since the criteria about maximum speed stands only if the agencies are real persons rather than advertising companies or any other non-individual based accounts. Considering the complexity of non-individual based activities as well as the differences in behavioral characteristics between individual and non-individual accounts, we decide to only focus on individual based accounts in our study.

After scrutinizing our initial result, we noticed that individual based accounts usually post tweets using official Twitter apps as well as third-party apps designed for personal uses, whereas non-individual accounts often adopt the customized app/platforms which provide professional tweeting functions such as group tweeting, event triggering or location modifications. At this point, since “source” property, indicating the client in use for tweeting, can imply (a) the type of device in use (e.g., iPhone, Android, iPad, Windows Phone, Blackberry, etc.), (b) official apps developed by Twitter or a third-party (e.g., Foursquare, Instagram, Path, Flickr, etc.), (c) the general usage and designed purpose of the app, we can use the “source” property to differentiate the individual based accounts from the non-individuals. Thus, we can successfully find out the
individual based accounts through parsing the “source” property by NLP (Natural Language Processing) toolkits.

Have we neglected any suspicious spoofing with a lower speed than 1,000 km/h? Moreover, the baseline cannot differentiate which tweet of an identified pair of geo-tagged tweets on earth has been falsified. Bearing all these questions in mind, in the next section, we attempt to optimize the baseline approach and describe our research methodology in details.

4 Research methodology

In this section, we describe our research methodology in details, which rests upon a geographic approach. The geographic features of this approach are embodied in the geographic facts and principals in use, such as the speed limit of commercial aircrafts and the shortest flights all over the world, the fastest ground vehicles in different countries, and the possibility to post tweets while being on board. In addition, we also rely on Tobler’s first law of geography (TFL) – the further the distance from the given geo-tag to the daily activity area, the larger the probability the given geo-tag has been spoofed – to determine the counterfeit location of an identified pair.

4.1 Identifying the pairs containing counterfeit locations

In addition to the constraints on maximum speed of travel, we also aim to consider other possible geographic facts for improving our ability to identify location spoofing. Let us consider the tweeting behavior of an average Twitter user travelling by air. Increasingly, more and more commercial airlines provide in-flight Wi-Fi access. According to a report from USA Today (Mutzabaugh, 2013), 38 percent of domestic flights are enabled for in-flight Wi-Fi. So we can infer that tweeting becomes possible for some of the flights. Even so, posting geo-tagged tweets is still a hard task mainly due to GPS signals inside the main cabin is weak and intermittent to some extent. Today, microwave-absorbing materials have been widely adopted in making airplane bodies because of its lightweight, high strength as well as perfect property of electromagnetic wave absorption (Chung, 2001). Since GPS signal is a typical electromagnetic wave, only if putting the smartphones near to an opened window, it is almost impossible for most
circumstances to receive GPS signals. Therefore, we do not consider the possibility to send geo-tagged tweets on board, so:

(i) If posting geo-tagged tweets on board is not considered, any fight passenger/potential twitter user $u$ can post geo-tagged tweet only if the plane is on the ground. Accordingly, we can confidently infer that, given $u$ has taken a fight during the time interval of posting a pair of tweets $t(u)$. $t(u)$ establishes only if $u$ travels longer than the shortest fight both in the distance $D_{a_{\text{min}}}$ and time duration $T_{a_{\text{min}}}$.

(ii) If $u$ tweets right before taking off and after landing, ideally, the average speed $v(u)$ should be almost as high as the average speed of that airplane $V_a$. If $u$ does not tweet immediately but wait for a while, $v(u)$ must be any value ranging from 0 to $V_a$.

(iii) There is a speed gap between the fastest aircraft $V_{a_{\text{max}}}$ and the fastest ground vehicle $V_{t_{\text{max}}}$.

![Figure 2 A d-d plot to show the portion of location spoofing](image)

Furthermore, we plot the distance and duration properties as a tuple $[\text{dist}(u), \text{dur}(u)]$ to a distance-duration plot (aka d-d plot, y-axis denotes distance, while y-axis indicates time duration.)
See Figure 2), all the pair of tweets with counterfeit location $S_{cf}$ should locate in the upper-left area (the dashed-line filled area) to the line $V_{max}$, as well as the region within the boundaries made up by $V_{d_{max}}$, $D_{a_{min}}$, $T_{a_{min}}$ and $V_{t_{max}}$.

The fastest speed of ground vehicle $V_{t_{max}}$ varies in different countries and regions of world. To be more specific, $V_{t_{max}}$ equals to 240 km/h in North America (the maximum civil operating speed of Amtrak), 320 km/h in Europe (the maximum operating speed of TGV in France, ICE 3 in Germany), 320 km/h in Japan (the maximum operating speed of Shinkansen, literally meaning new trunk line) and 380 km/h in China (the maximum operational speed of Harmony CRH 380A). Considering most of the countries do not have such fast trains, we use 320km/h as the default value of $V_{t_{max}}$ in the rest of countries or regions in the world. So $V_{t_{max}}$ is a list of values, varies according where the pair of tweets $t(u)$ locates. Moreover, Star Alliance, as one of the leading global airline network, has the largest number of daily departures more than 21,900. Using their timetable\(^2\), we found that $D_{a_{min}}$ is around 120 km as well as $T_{a_{min}}$ is roughly 45 minutes. Therefore, having all the necessary values, we can single out the eligible location spoofing pairs.

### 4.2 Determining the counterfeit location out of a pair

Although more solid geographic facts do facilitate us to effectively pick out reliable location spoofing pairs, the very counterfeit location in a pair is still unclear. We plan to rely on the Tobler’s First Law of Geography (TFL) to automatically identify those counterfeit locations. Since most humans usually live, work, commute and send tweets within a daily activity area (Sherman, 2005), the further the distance from the given geo-tag to the daily activity area, the larger the probability the given geo-tag has been spoofed (See Figure 3). Accordingly, we convert the task to find the counterfeit location out of a location spoofing pair to another task to check which one is further to the activity area. Moreover, the activity area can be articulated by a few self-generated geo-tagged tweets (the white dots in Figure 3). And in order to determine the

---

very counterfeit location of a pair, we compare the distances from both suspicious counterfeit locations of a pair to the center of the activity space, and assume the one with a further distance as the counterfeit location.

Figure 3 The location spoofing pair and the daily activity area
Of course, exceptions do exist. It is reasonable for the counterfeit locations to mix with other geo-tags in the same daily activity area. Under this circumstance, the former criteria might be less effective in determining the counterfeit location. At this point, we have to resort to a variety of geographic experiences as well as some supplementary clues in user profiles, contents of social media feeds and social relations (Cheng, 2013) to implement the identification.

Considering the existence of exceptions, we plan to find out how much percentage our approach can estimate the critical counterfeit location out of a pair. For doing this, we start with processing a certain amount of location spoofing pairs using the proposed approach, and then we process the same set of pairs by manual identifications. Through comparing their results, we can roughly determine what proportion of location spoofings can be automatically identified.

5 Results and discussions
In this section, we report the results of an empirical study of identifying location spoofings from millions of geo-tagged tweets. And the goal of this empirical study is two-fold: (a) to evaluate the effectiveness of our approach; (b) and to provide our understanding of this complex phenomenon.
### 5.1 Results

Following the procedure of our approach, we put the eligible tweet pairs sent by individual based accounts onto a d-d plot (see the overview map on the upper right corner of Figure 4), while the tweet pairs with counterfeit locations should locate in the upper-left area to the line (indicating $V_{a_{max}}$, 1000 per hour) as well as the region within the boundaries made up by $V_{a_{max}}$ (1000 km/h), $Da_{min}$ (120 km), $Ta_{min}$ (45 minutes) and $V_{t_{max}}$ (380 km/h in China, 240 km/h in U.S., 320 km/h in Europe, Japan and the rest part of the world). For example, the tweet pairs with counterfeit locations in U.S. lie in the dash line portion of Figure 4. In this way, we get a great number of pairs of tweets which containing at least one counterfeit geo-tags. For each pair, we determine the counterfeit geo-tag out of the pair through the automatic screening method (stated in Section 4.2). Moreover, to evaluate the accuracy, we designed a control experiment to process 1,000 random pairs through the manual identification. Comparing their results, we found more than 60 percent of both of their results are identical.

*Figure 4 Millions of tweets in the d-d plot*
Table 2 compares the proposed approach with the baseline. As shown, the result is categorized by “source” property (only the top five are listed). Obviously, the proposed approach is capable of identifying as double as the number of results by the baseline. In general, tweets with location spoofing account for 0.2 percent of the entire sample. Interestingly, we notice that the percentage of the pairs is much higher (around an order of magnitude) from the third-party apps (Foursquare, Instagram etc.) than the official twitter apps.

<table>
<thead>
<tr>
<th>Source</th>
<th># Tweets</th>
<th>Baseline</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Tweets</td>
<td># Pair</td>
<td>Pct.</td>
</tr>
<tr>
<td>Twitter for iPhone</td>
<td>1,623,069</td>
<td>480</td>
<td>0.03</td>
</tr>
<tr>
<td>Foursquare</td>
<td>116,416</td>
<td>769</td>
<td>0.66</td>
</tr>
<tr>
<td>Twitter for Android (phone)</td>
<td>781,123</td>
<td>341</td>
<td>0.04</td>
</tr>
<tr>
<td>Instagram</td>
<td>52,661</td>
<td>344</td>
<td>0.65</td>
</tr>
<tr>
<td>Twitter for Android (tablet)</td>
<td>95,872</td>
<td>58</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2,806,516</td>
<td>2,153</td>
<td>0.08</td>
</tr>
</tbody>
</table>

In Figure 5, we map out all the pairs at both global and city level (take New York City as an example) to analyze the spatiality of these counterfeit locations. Globally speaking, consistent with the global distribution of the entire sample (see Figure 1), the identified location spoofings are also unevenly located. More specifically, users in North America and West Europe select a variety of clients when sending out counterfeit locations, whereas Japanese and Southeast Asia location-spoofers prefer Android, while Brazilian location-spoofers prefer Foursquare. Moreover, we also observed curves with different lengths tying up the counterfeit location and another geo-tag possibly in the daily activity area. At both global and city level, the curves are of different length, that is to say, there is no explicit behavior pattern in space-time for location spoofing, which actually can be either thousands of miles apart or just within several meters. The location spoofing phenomena do exist in different scales.
Overall, the result shows that the detected location spoofing exists in different scales ranging from global to individual level. However, the identified evidences, only accounting for a small portion around 0.2 percent of the entire sample, does not dramatically affect the data quality of social media at the collective level. Nonetheless, we notice that location spoofing, as an emerging social media phenomenon, is not trivial or insignificant, especially when examining these evidences at the individual level. Hence, in the following subsection, we attempt to further contribute our understanding of this complex phenomenon.

5.2 Discussions

In general, we believe that at least the following three aspects warrant further discussions:

*Multiple motivations to create counterfeit locations*

For individual users, we notice that the act of creating counterfeit locations actually stems from various motivations. The use of location spoofing can effectively mislead the forger’s followers, social media service providers or even the government/military agencies, thereby protecting the geo-privacy. Moreover, the ability to access a special location can somehow reflect the privilege,
vocation or other identities of an individual. For example, only those military personnel are admitted into the military zones, while only the rich families are able to live in the upscale communities. Accordingly, spoofing a location to a special place has become an effective way to forge the locational identity of a social media user. The audience might mistakenly believe this location forger has the same identity as the ones who actually live, act or work at that location.

Tolerance to location spoofing

We notice that most third-party apps purposely tolerate some degree of location spoofing as a subtle form of business advertising or protecting individual geo-privacy. Take the commonly used function “check-in” as an example, it enables assigning the coordinates of an actual place (e.g., restaurant, cinema, supermarket) to an information entity. Once “checked-in” at any nearby candidate location (determined by the apps) other than the actual location, the locational information is actually spoofed. Indeed by doing this, while spurring the local business to offer offline services (e.g., coupons, advertisings) to social media users who have virtually been there, this location based function also encourages to check in more virtual places of local business that provide offline promotions.

Moreover, most social media allege to protect individual geo-privacy using location spoofing, also named as “location obfuscation” (Duckham and Kulik, 2005). Technically speaking, location obfuscation degrades the quality of locational information for manipulating, visualizing, sharing and other immediate functions that may expose the actual location. Even so, the original locational information is still stored in the server in order to support location-based services that requiring the actual locations of consumers. Undeniably, it has been a severe threat to store our actual locational information in a remote server, especially when we passively and continuously receive annoying location based advertising but do not know how and for what particular purposes our information has been used.

Uncertainty of locational information

The geo-tag, though coded in precise coordinates, might contain some inevitable built-in uncertainties. To unveil this uncertain nature, we attempt to investigate the whole process of
“tweeting”, which is composed of three major steps -- creating, composing and finally sending out a tweet. Usually, the spatial and temporal properties of a tweet do not describe the same space-time slice of the whole tweeting process, especially when the user posts the given tweet while travelling as well as costing some time to compose the given tweet. The inevitable time lapse as well as the space displacement largely represents the inherent uncertainty of tweeting as a virtual behavior in time-space.

Indeed, the uncertainty of locational information, similar to any other form of geographic data, “is built into the knowledge production process, and not just because of human limitation” (Couclelis, 2003, 173). Furthermore, the uncertainty of geographic information often ignored (Goodchild, 1998). As discussed, the uncertainty of locational information in location based big data might invoke spoofings to some extent. In order to mitigate the interferences by uncertainties, we can model the locational information by fuzzy set (Altman, 1994; Robinson, 2003), describe by detailed metadata (Comber et al, 2006), or visualized by cartographic techniques, such as blurring the feature, fading colors to distinguish the uncertain area, or using a round area to indicate the possibility of appearance within that given area rather than at a single position (Goodchild, 1998).

6 Summary and conclusions

This paper has proposed a geographic approach to detect location spoofings in big data. The geographic features of this approach are embodied in the geographic facts (e.g., the maximum speed for commercial aircrafts and the shortest flights all over the world, the fastest ground vehicles in different countries) and the geographic principles (e.g., TFL) in use. After applying this approach to an empirical study with millions of tweets, we found that this approach is capable of effectively detecting certain types of location spoofings. Specifically, the identified location spoofings are spatially distributed at different scales, ranging from global to individual level. Since the total number of location spoofing accounts for a trivial portion around 0.2 percent of the entire sample, it does little impact on the data quality of geo-enabled social media, which is one of major sources of big data. But admittedly, for those counterfeit locations
continuously sent out at an identical spot or with a speed lower than the maximum speed of the
ground vehicles, our approach is not applicable. In fact, it is not our goal to develop an approach
to detect all kinds of location spoofings in big data - an extremely challenging (if not entirely
impossible) task. Nonetheless, we argue that the proposed approach is still of significance in
terms of automatically filtering large-scale geo-tagged social media feeds or any other types of
location based big data, evaluating the utility of geographic facts in identifying location spoofing,
and suggesting us possible directions for further improvements, such as to integrate with more
targeted geographic knowledge, or to utilize the social hierarchical and crowd-sourcing
identification.
Ultimately, location spoofing is not an issue fundamentally different from other deceptive
aspects of big data or even the real world. Indeed, location spoofing has been (ab)used by
government/military agencies, big corporations, and average people for multiple purposes.
Contingent upon the specific circumstances, location spoofing could have significant impacts on
our daily lives, social developments, and even national security. From a strict technical
perspective, location spoofing, at least those caused by uncertainties, is inevitable (Couclelis,
2003). Based on the above consideration, rather than trying to “control” such a complicated issue,
a reasonable countermeasure is to directly “confront” it. Actually, one primary purpose of this
paper is to alert big data users, generators, service providers and regulators to confront the
location spoofing phenomena in big data. Only then can we holistically understand diverse kinds
of location spoofings, the possible motivations, and the inherent uncertain nature. In this sense,
to successfully identify, classify and analyze the location spoofing has moved a step forward.
Though still at a preliminary stage, we hope this paper raises the awareness on the multiple
dimensions of the location spoofing issue whenever location based big data are used in research
or policy-making.
Reference:
spatial analysis. *International Journal of Geographical Information Systems, 8*(3),


topics in security, Washington, DC.


applications. 60-64. Napa Valley, California.


Accounting for Heteroscedasticity in Big Data

Francisco Nibau Antunes†, Aidan O’Sullivan‡, Filipe Rodrigues†, Francisco Pereira†

mail to: camara@smart.mit.edu

†Center of Informatics and Systems of the Univ. of Coimbra, Portugal
‡Singapore-MIT Alliance for Research and Technology, Singapore

July 15, 2014

Abstract

For regression problems, the general practice is to consider a constant variance of the error term across all data. This aims to simplify an often complicated model and relies on the assumption that this error is independent of the input variables. This property is known as homoscedasticity. On the other hand, in the real world, this is often a naive assumption, as we are rarely able to exhaustively include all true explanatory variables for a regression. While Big Data is bringing new opportunities for regression applications, ignoring this limitation may lead to biased estimators and inaccurate confidence and prediction intervals.

This paper aims to study the treatment of non-constant variance in regression models, also known as heteroscedasticity. We apply two methodologies: integration of conditional variance within the regression model itself; treat the regression model as a black box and use a meta-model that analyzes the error separately. We will compare the performance of both approaches using two heteroscedastic datasets.

Although accounting for heteroscedasticity in data increases the complexity of the models used, we show that it can greatly improve the quality of the predictions, and more importantly, it can provide a proper notion of uncertainty or “confidence” associated with those predictions. We also discuss the feasibility of the solutions in a Big Data context.

Keywords Heteroscedasticity, Gaussian Processes, Quantile Regression
1 Introduction

Since the beginning of the 21st century, alongside the IT revolution, we have been witnessing a deluge of data, at some point referred to as “Big Data”: a very large, complex and heterogeneous collection of information, either structured or unstructured, which increases at a high rate over time. One of the main reasons for this data explosion is the proliferation of ubiquitous sensing-systems. For example, nowadays, even the simplest smartphone is usually equipped with, at least, GPS and accelerometer sensors. Similarly, many cities already have sensors gathering information from traffic, weather, security cameras, emergency systems, and so forth, thus making dataset sizes reach unprecedented levels.

For practical reasons, data is often treated as having more signal than noise, and a well behaved noise structure. This structure usually relies on assuming non-biased models and constant variance, typically formulated as a “white noise” Gaussian distribution, \( \mathcal{N}(0, \sigma^2) \). The problem is that, more often than not, reality is not as “well behaved” and such assumptions may become unrealistic and inappropriate. Consider the problem of travel time prediction using a few typical variables (e.g. time of day, day of week, latest observed travel times from sensors, flow data from sensors, weather status). For many urban areas, assuming a homoscedastic noise model amounts to saying that the travel time only depends on those specific variables. This assumption is seldom realistic, since in practice there exist fluctuations due to a myriad of phenomena (excess/lack of demand, traffic incidents, special events, errors in sensors). Indeed, under certain circumstances (e.g. heavy rain, working day), the travel times inherently vary much more than others (e.g. clear skies, weekend night) and the system is too complex to efficiently capture all components.

In this paper we demonstrate the critical importance of taking into account the heteroscedasticity in the data, by comparing the performance of homoscedastic and heteroscedastic approaches on different problems. We show the problems that arise when the assumption of constant variance is violated, and we propose some guidelines for the treatment of heteroscedasticity.

Mathematical approaches for this problem range from the classic statistical models to the more recent machine learning techniques. In a very general way, the majority of the regression models are of the form

\[
y = f(x) + \varepsilon,
\]

which establishes a relationship between the output variable we want to study \( y \), with other (input) variables \( x = [x^1, \ldots, x^D]^{\top} \), where \( D \) denotes the number of explanatory variables, by means of a certain function \( f \). We usually rewrite \( f \) as \( f(x, w) \), where \( w = [w_1, \ldots, w_D]^{\top} \) are the model parameters. In this kind of models, \( \varepsilon \) is the error or residual, which is usually assumed to be a random variable, with mean zero and some variance. Whenever this variance is constant, the model is considered homoscedastic, if it is assumed to vary with \( x \), it is considered heteroscedastic. The most common option is to assume a homoscedastic variance, with \( \varepsilon \sim \mathcal{N}(0, \sigma^2) \). Most of these models appear in the form of regression/classification or time series analysis.

Although it may be acceptable to expect a roughly unbiased model (i.e. error mean = 0), it is often unrealistic to assume constant variance. A possible way to go around this issue
is to consider heteroscedastic models. Assuming variable error along our input variables $x$ should result in better predictions and more accurate confidence intervals.

While the improvements in accuracy may be negligible with respect to the mean (if the homoscedastic model has little bias), we have much to gain in terms of confidence intervals. For example, from the user’s point of view, it is completely different to know that the bus arrival time will be 10 minutes with a confidence interval of 10 minutes, or with a confidence interval of just 1 minute.

This paper is organized as follows: Section 2 discusses general regression methods and the treatment of heteroscedasticity with emphasis on linear models. In Section 3, we move the discussion to more powerful non-linear regression models: Gaussian Processes. Section 4 presents quantile regression approaches for the problem of non-constant variances. We compare the different approaches in Section 5 using different datasets, and we present the conclusions in Section 6.

## 2 Linear models for regression

We now review the traditional Multiple Linear Regression (MLR) model and its problems with heteroscedasticity. This method has been widely studied and it continues to be a standard on statistical analysis in the majority of fields, mainly because of its interpretation simplicity and computational speed. The MLR model is defined as:

$$
y_i = w_0 + w_1 x_{i1} + w_2 x_{i2} + ... + w_D x_{iD} + \varepsilon_i, \ i \in \{1, ..., N\},$$

or in matrix notation,

$$
y = f(X, w) + \varepsilon = X^\top w + \varepsilon,$$

where $y = [y_1, y_2, ..., y_N]^\top$, $X$ is called the design matrix, a $N \times (D + 1)$ matrix such that each column $j \in \{1, ..., D\}$ is $x_j = [1, x_{1j}, x_{2j}, ..., x_{Nj}]^\top$, and $\varepsilon = [\varepsilon_1, \varepsilon_2, ..., \varepsilon_N]^\top$. So, the MLR states that $y$ is a weighted linear combination of the input $x$. Alongside a few more assumptions stated by the Gauss-Markov Theorem [5], $\varepsilon$ is assumed to follow the standard Gaussian distribution with fixed variance $\sigma^2$, that is, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Under this assumption, maximizing the (log) likelihood function w.r.t to the parameters $w$ is equivalent to minimizing the least-squares function [1], that is:

$$
\hat{w} = \arg \max_w \sum_{i=1}^{N} \log \mathcal{N}(y_i | x_i^\top w, \sigma^2 I) = \arg \min_w \sum_{i=1}^{N} (y_i - w^\top x_i)^2.
$$

The solution of (3) is called the Ordinary Least-Squared (OLS) estimator and it is given by

$$
\hat{w} = (XX^\top)^{-1} X y.
$$
This estimator has the following properties for the mean and variance:

\[
E[\hat{\mathbf{w}} | \mathbf{X}] = E[(\mathbf{X} \mathbf{X}^T)^{-1} \mathbf{X} \mathbf{y}] = E[(\mathbf{X} \mathbf{X}^T)^{-1} \mathbf{X} \mathbf{X}^T \mathbf{w} + \mathbf{e}] \\
= E[(\mathbf{X}^T)^{-1} \mathbf{X}^T \mathbf{X}^T \mathbf{w} + \mathbf{e}] = E[\mathbf{w}] = \mathbf{w}
\]  

(5)

\[
\nabla \hat{\mathbf{w}} | \mathbf{X} = (\mathbf{X} \mathbf{X}^T)^{-1} \mathbf{X} \mathbf{X}^T \mathbf{X} \mathbf{X}^T)^{-1},
\]

(6)

where \( \Phi \) is a diagonal matrix with \( \Phi_{ii} = \nabla \mathbf{e}_i = \sigma^2 \). Since the error term is homoscedastic, i.e., \( \nabla \mathbf{e}_i = \sigma^2 \) for all \( i \), then \( \Phi = \sigma^2 \mathbf{I} \), where \( \mathbf{I} \) is the identity matrix. So (7) can be simplified to

\[
\nabla \hat{\mathbf{w}} | \mathbf{X} = \sigma^2 (\mathbf{X} \mathbf{X}^T)^{-1}.
\]

(8)

Under the assumptions of the Gauss-Markov theorem, (4) is the best linear unbiased estimator of the covariance matrix of \( \mathbf{w} \). From (5) and (7) we can see that \( \hat{\mathbf{w}} \) is an unbiased estimator and its variance is influenced by the error variance. Notice that, under the presence of heteroscedastic residuals, \( \hat{\mathbf{w}} \) will stay untouched, in contrast to its variance, which will be a function of \( \sigma^2 \). Hence, the associated tests of significance and confidence intervals for the predictions will no longer be effective. In [23] the reader can find a brief but practical guide to some of the most common linear model assumption violations, including the usual assumption of homoscedastic residuals.

To overcome this problem, White [32] suggested a heteroscedasticity-consistent covariance matrix. In its simplest form, this amounts to setting \( \Phi \) to be a \( N \times N \) diagonal matrix whose elements are \( \Phi_{ii} = E(\mathbf{e}_i^2) = \nabla \mathbf{e}_i \). Here the problem is that we do not know the form of \( \nabla \mathbf{e}_i \), so we need to estimate it. Under certain conditions, this can be achieved by constructing the consistent estimator

\[
\hat{\nabla} \mathbf{e}_i = \frac{1}{N} \sum_{i=1}^{N} \mathbf{e}_i^2 \mathbf{x}_i \mathbf{x}_i^T
\]

(9)

However, \( \mathbf{e}_i^2 \) it is not observable. Fortunately, it can be estimated by

\[
\hat{\mathbf{e}}_i^2 = y_i - \mathbf{x}_i^T \hat{\mathbf{w}}
\]

(10)

thus leading us to consider the following estimator for \( (\mathbf{X} \Phi \mathbf{X}^T / N) \),

\[
\frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{e}}_i^2 \mathbf{x}_i \mathbf{x}_i^T.
\]

(11)

White proved that (11) is a heteroscedastic-consistent covariance matrix estimator of the “asymptotic covariance matrix” given by (7). For an extensive detailed reading on the subject besides [32], we suggest, [22], [21], [27] and the more recent [19], [33] and [20].

The MLR can be extended to incorporate linear combinations of non-linear functions of the input, thus allowing \( y \) to be a non-linear function of \( \mathbf{x} \), but still a linear function of the weights \( \mathbf{w} \). Despite this generalization, the demand for pure non-linear models continued.
Specially concerning the treatment of non-constant variance, many non-linear heteroscedastic models have been presented since the beginning of the 80s, mainly applied to finance data [7], where time-dependent volatility takes its upmost form. There were suggested models like the ARCH, ARMA-CH and its variations [4], [30], [34] and, more recently, regarding regression by Gaussian Processes [26], there have been developments, with the introduction of the Heteroscedastic Gaussian Process concept [10], [12], [15], which models the error term with another Gaussian Process dependent on the input (see Section 3).

Causes for heteroscedasticity vary from case to case, but most of them are related to the model misspecification, measurement errors, sub-population heterogeneity, noise level or it is just an natural intrinsic property of the dataset. To determine the application feasibility of heteroscedastic models, we should first test its presence. There is a number of statistical tests that allow, up to some extent, checking the presence of heteroscedasticity in data, this includes Breuch-Pagan [2], Cook-Weisberg [6], White [32] and Goldfeld-Quandt [11] tests, amongst others. Other less statistically significant test, although very practical, is the visual inspection of the model residuals. The scatter-plot of the standardized residuals against the predicted values of \(y\) should lie uniformly around the zero horizontal line. Otherwise, heteroscedasticity should be suspected.

### 3 Gaussian Process Regression

Gaussian Processes (GP) propose a very different angle to regression than MLR, allowing for modeling non-linear relationships and online learning. Their non-parametric Bayesian nature gives them a sufficiently good flexibility to be used in a vast variety of regression and classification problems. In the following sections we will present the standard GP model and an heteroscedastic variant.

#### 3.1 Standard GP

As in [26], a Gaussian Process is a (possibly infinite) collection of random variables, \(f = (f_t, t \in T)\), any finite number of which have a joint Gaussian distribution. \(T\) is a set of indexes and if it is finite, the GP reduces to joint Gaussian distribution. Each GP is completely determined by its mean and covariance (or kernel) functions. These functions, respectively \(m_f(x)\) and \(k_f(x,x')\), can be defined as:

\[
\begin{align*}
m_f(x) &= \mathbb{E}[f(x)] \quad (12) \\
k_f(x,x') &= \text{Cov}(f(x), f(x')) = \mathbb{E}[(f(x) - m_f(x))(f(x') - m_f(x'))]. \quad (13)
\end{align*}
\]

Thus we can simply denote a Gaussian Process as \(\mathcal{GP}(m_f(x), k_f(x,x'))\). In order to keep things simple, \(m_f(x)\) is often considered zero (or constant). This is a common practice and it is not a very restrictive one, since the mean of the posterior process is not confined to the mean function of the GP.

The standard Gaussian Process regression is a Bayesian framework which assumes a GP
prior over functions, i.e

\[ y = f(x) + \epsilon, \]

(14)

where \( \epsilon \sim \mathcal{N}(0, \sigma^2) \) and \( f(x) \sim \mathcal{GP}(m_f(x), k_f(x, x')) \).

Assuming \( m_f(x) = 0 \), the prior over the latent function values is then given by:

\[ p(f|x_1, x_2, ..., x_n) = \mathcal{N}(0, K_f), \]

(15)

where \( f = [f_1, f_2, ..., f_n]^\top \), \( f_i = f(x_i) \) and \( K_f \) is the covariance matrix, with its elements given by \( [K_f]_{ij} = k_f(x_i, x_j) \). Many forms for the covariance functions can be considered, each one typically having a number of free hyper-parameters, which we refer to as \( \theta \). For a particular application, we need to fix, a priori, a family of covariance functions and optimize the kernel w.r.t. the hyper-parameters. One way of setting the hyper-parameters is to maximize the marginal likelihood given by

\[ \log(y|X, \theta) = -\frac{1}{2} y^\top K_y^{-1} y - \frac{1}{2} \log |K_y| - \frac{N}{2} \log(2\pi), \]

(16)

where \( K_y = K_f + \sigma^2 I \) and \( K_f \) are, respectively, the covariance matrix for the noisy targets \( y \) and noise-free latent variable \( f \).

Having set the covariance function and its corresponding hyper-parameters, the conditional distribution of a new test point \( x_* \) is given by:

\[ f_*|X, y, x_* \sim \mathcal{N}(\mathbb{E}[f_*], \mathbb{V}[f_*]), \]

(17)

with

\[ \mathbb{E}[f_*] := \mathbb{E}[f_*|X, y, x_*] = k_{f_*}^\top[K_f + \sigma^2 I]^{-1} y, \]

(18)

\[ \mathbb{V}[f_*] = k_{f_*^*} - k_{f_*}^\top[K_f + \sigma^2 I]^{-1} k_{f_*}, \]

(19)

where we introduced the notation \( k_{f_*} = k_f(X, x_*) \) and \( k_{f_*^*} = k_f(x_*, x_*) \).

Thus we can construct a \((1 - \alpha)\)% Confidence Interval (CI) for \( y_* \) as follows

\[ k_f(x_*, X)[K_f + \sigma^2 I]^{-1} y \pm z_{\frac{\alpha}{2}} \sqrt{k(x_*, x_*) - k_f(x_*, X)^\top(K_f + \sigma^2 I)^{-1} k_f(X, x_*)}, \]

(20)

where \( z_{\frac{\alpha}{2}} \) is the \( \frac{\alpha}{2} \)th order quantile of \( \mathcal{N}(0, 1) \).

This allows us to construct confidence intervals under the constant variance assumption. In the next section we extend this model in order to relax this assumption.

### 3.2 Heteroscedastic GP

In this section we will closely follow the work of Gredilla et al, [12]. As seen in the previous section, the standard GPs assume a constant variance, \( \sigma^2 \), throughout the input, \( x \). We will now relax this assumption.
To define a Heteroscedastic GP (HGP), besides placing a GP prior on \( f(x) \) as in (14), we also place a GP prior on the error term, so that we have:

\[
y = f(x) + \varepsilon,
\]

with \( f(x) \sim \mathcal{GP}(0, k_f(x, x')) \) and \( \varepsilon \sim \mathcal{N}(0, r(x)) \), where \( r(x) \) is an unknown function. To ensure positivity and without losing generality, we can define \( r(x) = e^{g(x)} \), where \( g(x) \sim \mathcal{GP} (\mu_0, k_g(x, x')) \).

After fixing both covariance functions, \( k_f \) and \( k_g \), the HGP is fully specified and depends only on its hyper-parameters. Unfortunately, exact inference in the HGP is no longer tractable. To overcome this issue, the authors propose a variational inference algorithm which establishes a Marginalized Variational (MV) bound, for the likelihood function, given by:

\[
F(\mu, \Sigma) = \log \mathcal{N}(y|0, K_f + R) - \frac{1}{4} tr(\Sigma) - \text{KL}(\mathcal{N}(g|\mu, \Sigma) || \mathcal{N}(g|\mu_01, K_g)),
\]

where \( R \) is a diagonal matrix such that \( R_{ii} = e^{\mu_0 - \frac{1}{2} \Sigma_{ii}} \), \( K_g \) is the covariance matrix resulting from the evaluation of \( k_g(x, x') \), \( \mu \) and \( \Sigma \) are the parameters from the variational distribution \( q(g) = \mathcal{N}(g|\mu, \Sigma) \), and KL is the Kullback-Leibler divergence. The hyper-parameters of this Variational Heteroscedastic GP (VHGP) are then learned by setting the following stationary equations:

\[
\frac{\partial F(\mu, \Sigma)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial F(\mu, \Sigma)}{\partial \Sigma} = 0.
\]

The solution for this system is a local or global maximum and it is given by:

\[
\mu = K_g(\Lambda - \frac{1}{2} I) + \mu_01 \quad \text{and} \quad \Sigma^{-1} = K_g^{-1} + \Lambda,
\]

for some positive semidefinite diagonal matrix \( \Lambda \). We can see that both \( \mu \) and \( \Sigma \) depend on \( \Lambda \), letting us rewrite the MV as a function of \( \Lambda \)

\[
F(\mu(\Lambda), \Sigma(\Lambda)) = F(\Lambda),
\]

which needs to be maximized w.r.t. to the \( N \) variational parameters in \( \Sigma \). At the same time, it is possible to maximize \( F \) w.r.t to the model hyper-parameters \( \theta \). After learning the variational parameters and model hyper-parameters, \( (\Lambda^*, \theta^*) \), by maximizing \( F \), we need to approximate the predictive distribution \( p(y_*|x_*, X, y) \) for a new test input-output point \( (x_*, y_*) \), which again is intractable. So, within the variational approach, the predictive distribution of \( y_* \) can be approximated by:

\[
q(y_*) = \int \mathcal{N}(y_*|a_*, c_*^2 + e^{\theta^*}) \mathcal{N}(g_*|\mu_*, \sigma_*^2) dg_*,
\]

with \( a_* = k_{f*}^T(K_f + R)^{-1}y, c_*^2 = k_{f*} - k_{f*}^T(K_f + R)^{-1}k_{f*}, \mu_* = k_{g*}^T(\Lambda - \frac{1}{2} I)1 + \mu_0 \) and \( \sigma_*^2 = k_{g*} - k_{g*}^T(K_g + \Lambda^{-1})^{-1}k_{g*} \). For a more detailed derivation, please refer to [12]. For an
introduction on the dense theory on the variational approximation framework, we recommend [31] and [9].

Although (27) is not analytically tractable, its mean and variance can be computed exactly:

\[
\mathbb{E}_q[y_*|x_*, X] = a_* \quad (28)
\]
\[
\mathbb{V}_q[y_*|x_*, X] = c_*^2 + e^{\mu_*} + \sigma_*^2. \quad (29)
\]

We are then able to construct the following \((1 - \alpha)\)% heteroscedasticity-consistent CI for \(y_*\):

\[
a_* \pm z_{\frac{\alpha}{2}} \sqrt{c_*^2 + e^{\mu_*} + \sigma_*^2}, \quad (30)
\]

with \(\alpha \in ]0, 1[.\)

### 3.3 GPs for Big Data

The biggest drawback of Gaussian Processes is that they suffer from computational intractability issues for very large datasets. As seen in [26], typically the complexity scales around \(O(N^3)\) so, for \(N > 10000\), both storing and inverting the covariance matrix proves to be a prohibitive task on most modern machines. More recently, and specially concerning the treatment of Big Data, [13] proposed a stochastic variational inference algorithm that is capable of handling millions of data points. By variationally decomposing the GP and making it only dependent on a specific set of relevant \(M\) inducing points, the authors lowered the complexity to \(O(M^3)\). Also concerning the application of GPs to large datasets, [28] suggested a new sparse approximation resulting from the combination of both global and local approaches. There is a vast literature regarding GP approximations. For an overview on some of these approximation methods we suggest [25].

### 4 Quantile Regression

Since, in some cases, integrating a heteroscedastic component in the model may not be practical due to overwhelming complexity or computational intractability, an alternative is to do post-model analysis. In this case, we analyze the output performance of a certain prediction process, whose internal details are not necessarily known or understood, and model its observed error. This is the approach proposed by [14] and [18], where the authors focus on time series prediction. In [24], we extended this approach with more complex Quantile Regression (QR) models applied to general predictors. Our algorithm treats the original prediction model as a black box and generates functions for the lower and upper quantiles, respectively providing the lower and upper bounds of the prediction interval. Despite this flexibility, being downstream to the model itself comes with a cost: this approach will hardly correct earlier modeling mistakes and will only uncover such limitations (by presenting very wide prediction bounds).
In contrast to least-squared based regression approaches, QR fits the regression parameters to the quantiles, instead of fitting them to the conditional mean. By doing so, this type of regression tends to be more robust against outliers and does not depend on the common assumptions of normality or symmetry of the error. Following [17] and [8], we can denote the conditional quantile function as

\[ Q_y(\tau|X), \]  

where \( \tau \) is the quantile order we want to estimate.

In its simplest approach, if \( Q_y \) has a parametric structure, i.e., if \( Q_y = \xi(X, w) \), to estimate the desired quantile \( \tau \), we proceed to the following optimization

\[ \min_w \sum_{i=1}^{N} \rho_\tau(y_i - \xi(X, w)), \]  

where \( \rho_\tau(.) \) is the tilted loss function represented in Figure 1.

As we can see, \( Q_y \) can have a variety of functional structures, either parametric or non-parametric. In particular, we can define the Gaussian Processes for Quantile Regression (GPQR), by placing a GP prior over \( Q_y \), that is,

\[ Q_y(\tau|X) \sim GP(0, k(X, X')) \]  

where \( k \) is the covariance function.

In this paper, we will use the meta-model perspective suggested by [24] and [8]. For that, consider a random black box model that generates predictions, \( \hat{y}_i \). We will focus on the error of such predictions in terms of the deviations \( d_i = y_i - \hat{y}_i \), where \( y_i \) is the series of the observed values. The aim of this model is to associate lower and upper bounds for the predicted target value, i.e., to determine the functions \( d^{\tau_-}(X) \) and \( d^{\tau_+}(X) \) which provide, respectively, the lower and upper quantiles, \( \tau_- \) and \( \tau_+ \), defined a priori.

When seeking for the quantile \( \tau \), the loss of choosing \( d^\tau \), instead of \( \tau \), is quantified by

\[ \rho_\tau(u) = u(\tau - 1_{u<0}), \]  

where \( 1_{u<0} \) is the indicator function.
with \( u = d - d^\tau \), which is equivalent to

\[
\rho_{\tau}(d - d^\tau) = \begin{cases} 
\tau(d^\tau - d), & d \geq d^\tau \\
(1 - \tau)(d - d^\tau), & d < d^\tau.
\end{cases}
\] (35)

Within the GPQR approach, the posterior distribution of the quantile function \( Q_y \), is

\[
p(Q_y|D, \theta) = \frac{1}{Z} p(D|Q_y, \theta)p(Q_y|\theta),
\] (36)

where \( \theta \) is the set of the covariance function hyper-parameters, \( Z \) a normalization factor and \( D = (X, y) \) the training set.

The hyper-parameters are estimated by maximizing the likelihood function,

\[
\arg \max_\theta p(Q_y|D, \theta),
\] (37)

which was proved to be equivalent to minimizing the tilted loss function. Hence, the likelihood becomes

\[
p(D|Q_y) = \left(\frac{\tau(1 - \tau)}{\sigma}\right)^N e^{-\sum_{i=1}^N \frac{u_i^2}{\sigma} (\tau_{1_{u_i < 0}})},
\] (38)

where \( \sigma \) is the standard deviation of the Asymmetric Laplace distribution and \( u_i = d_i - d^\tau_i \).

With this meta-model approach, we introduced a new kind of interval: the Prediction Interval (PI). Although the differences between both CI and PI are subtle, it is important to distinguish the two concepts. If on the one hand a \((1 - \alpha)\%\) CI is expected to contain a certain population parameter (e.g. mean, variance) in, at least, \((1 - \alpha)\%\) of the times, on the other hand, a prediction interval should cover the next predicted value, at least, \((1 - \alpha)\%\) of the times.

## 5 Experiments

In this section we will compare the performances of the GP, VHGP and GPQR approaches to the problem of the heteroscedasticity, using two one-dimensional datasets. From now on we will refer to the GPQR only by QR. More than assessing the pointwise quality of the predictions using the standard performance measures, we are mostly interested in evaluating how accurately the constructed CIs and QR prediction intervals handled the volatility in the data.

### 5.1 Performance measures

For the single-point predictions we will use the following standard and well accepted measures: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), RAE (Root Absolute Error), RRSE (Root Relative Squared Error) and COR (Pearson’s linear correlation coefficient).
For the interval quality evaluation we will consider some of the performance measures suggested in [24], although with slight modifications in order to extend their applicability to both kind of intervals, CI and PI. Let \( l_i \) and \( u_i \) be the series of the lower and upper bounds of a confidence/prediction interval for \( \hat{y}_i \), respectively. Consider also the series of the real target values \( y_i \). Then we can define the following measures:

- **Interval Coverage probability (ICP),**
  \[
  \text{ICP} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{y_i \in [l_i, u_i]},
  \tag{39}
  \]

- **Relative mean interval length (RMIL),**
  \[
  \text{RMIL} = \frac{1}{N} \sum_{i=1}^{N} \frac{(u_i - l_i)}{|y_i - \hat{y}_i|},
  \tag{40}
  \]

- **Coverage-length-based criterion (CLC2),**
  \[
  \text{CLC2} = e^{(-\text{RMIL}(\text{ICP} - \mu))},
  \tag{41}
  \]

  where \( \mu \) is a controlling parameter.

In our case, for both CI and QR bounds, ICP should be as close to \( (1 - \alpha) = (\tau^+ - \tau^-) = 0.95 \) as possible. RMIL intrinsically expresses the idea that, for a large observed error, we need to allow large intervals, so that the predicted value can be covered. It can be seen as a weighted average relative to the actual observed error. For CLC2, we can see that \( \mu \) represents the ideal value for ICP. When \( \mu = \text{ICP} \), CLC2 reaches its optimal value, 1. In the same exponent, RMIL can be regarded as an amplification factor of the “importance” of being close to the desired ICP. CLC2 is a simplification of CLC [16] that replaced an arbitrary weight, \( \eta \), for RMIL as just described.

### 5.2 Datasets

We used the motorcycle dataset from [29] composed of 133 data points and a synthetically generated toyset consisting of 1000 data points exhibiting a linear relationship with local instant volatility. For each dataset we ran three models: GP, VHGP and QR. The QR model used as input the predictions from the two others. For both GP and VHGP we generated predictions in a 10-fold cross-validation fashion, using the code freely available from [26] and [12]. For the QR, we used the code available from [3]. We fixed \( \alpha = 0.05 \), \( \tau^- = 0.025 \) and \( \tau^- = 0.975 \) so that we have equivalent interval coverages.
Table 1: Performance Measures Results Table

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>ICP</th>
<th>RMIL</th>
<th>CLC2</th>
<th>RMSE</th>
<th>MAE</th>
<th>RAE</th>
<th>RRSE</th>
<th>COR</th>
</tr>
</thead>
<tbody>
<tr>
<td>GP</td>
<td>motorcycle</td>
<td>0.933</td>
<td>29.508</td>
<td>1.684</td>
<td>23.343</td>
<td>17.525</td>
<td>0.454</td>
<td>0.485</td>
<td>0.875</td>
</tr>
<tr>
<td>VHGP</td>
<td>motorcycle</td>
<td>0.955</td>
<td>14.562</td>
<td>0.931</td>
<td>23.666</td>
<td>17.725</td>
<td>0.459</td>
<td>0.491</td>
<td>0.871</td>
</tr>
<tr>
<td>QR over GP</td>
<td>motorcycle</td>
<td>0.955</td>
<td>12.287</td>
<td>0.941</td>
<td>n.a</td>
<td>n.a</td>
<td>n.a</td>
<td>n.a</td>
<td>n.a</td>
</tr>
<tr>
<td>QR over VHGP</td>
<td>motorcycle</td>
<td>0.947</td>
<td>8.526</td>
<td>1.023</td>
<td>n.a</td>
<td>n.a</td>
<td>n.a</td>
<td>n.a</td>
<td>n.a</td>
</tr>
<tr>
<td>GP</td>
<td>toyset</td>
<td>0.931</td>
<td>329.750</td>
<td>525.978</td>
<td>253.431</td>
<td>152.363</td>
<td>0.202</td>
<td>0.285</td>
<td>0.958</td>
</tr>
<tr>
<td>VHGP</td>
<td>toyset</td>
<td>0.940</td>
<td>88.707</td>
<td>2.428</td>
<td>253.656</td>
<td>149.791</td>
<td>0.198</td>
<td>0.286</td>
<td>0.959</td>
</tr>
<tr>
<td>QR over GP</td>
<td>toyset</td>
<td>0.910</td>
<td>85.745</td>
<td>30.870</td>
<td>n.a</td>
<td>n.a</td>
<td>n.a</td>
<td>n.a</td>
<td>n.a</td>
</tr>
<tr>
<td>QR over VHGP</td>
<td>toyset</td>
<td>0.956</td>
<td>19.762</td>
<td>0.888</td>
<td>n.a</td>
<td>n.a</td>
<td>n.a</td>
<td>n.a</td>
<td>n.a</td>
</tr>
</tbody>
</table>

5.3 Results

Table 1 summarizes the overall obtained results. The variance evolution of the GP and VHGP are plotted in Figures 2 and 3. Figures 4-7 compare the three different approaches within each dataset.

In terms of interval coverage, we can see that all models performed well, although the VHGP and QR had an ICP sightly closer to the ideal value, except for the toyset. It is clear that the homoscedastic model (GP) was able to reach such ICP values because of its large, but constant, variance (see Figures 2 and 3). However, ICP cannot be regarded independently from the others. High values of ICP do not necessarily imply good intervals, they merely mean that the constant variance assumed for the model is sufficiently large for the CIs to coverage almost every data point. So, the GP did achieve, in fact, reasonable good marks for ICP, but it did it so at the expense of larger (and constant) interval ranges. As a result these values are best considered in combination with RMIL, which relates the interval bounds and the actual observed error. Unsurprisingly, the RMIL values for the GP are the larger ones, which means that, on average, the considered intervals were larger than required to cover the data or too narrow to handle its volatility. On the other hand, the heteroscedasticity-consistent approaches, VHGP and QR, were able to dynamically adapt their error variances to the data volatility, leading to tighter intervals that more accurately reflect the uncertainty in the data, as we can see from the values of RMIL and CLC2.

The results obtained for the toyset are consistent with and show a similar pattern to those of the motorcycle dataset. The values of RMIL and CLC2 obtained for the VHGP show a great improvement over the standard homoscedastic GP as observed in the toyset and as would be expected given the heteroscedasticity present in the data. Again, as observed in the toyset, the addition of a quantile regression meta-model approach to the standard homoscedastic GP results in much better performance in terms of RMIL and CLC2. The RMIL in particular is now comparable to the value obtained using VHGP. There is some reduction in ICP which is undesirable but we still obtain a high value, albeit slightly less than desired. The QR-VHGP combination again exhibits the best overall performance in comparison to the three previous methods.
Figure 2: GP versus VHGP variances in the motorcycle dataset.

Figure 3: GP versus VHGP variances in the toyset.
Figure 4: Comparison between GP and QR over GP predictions for the motorcycle dataset.

Figure 5: Comparison between VHGP and QR over VHGP predictions for the motorcycle dataset.
6 Conclusions and Future Work

In this paper we have explored the problem of heteroscedasticity for regression problems. Using data with this characteristic property, both real and synthetic, we have compared the performance of a GP model which has been augmented to handle heteroscedasticity with that of a standard GP. The results obtained highlight the improved performance of this more complex model and provide motivation for its use in such settings. Given the current popularity of GPs as an advanced machine learning tool for Big Data problems, this is an important result that demonstrates that while the GP framework is extremely flexible and
powerful, the problem of heteroscedasticity is an issue that must be considered and handled using appropriate tools. With this in mind we also studied the use of quantile regression as a post processing means of taking heteroscedasticity into account even when we cannot or don’t want to make structural changes in our model.

The results show that the post-processing approach can improve the results of a homoscedastic model to the point of being comparable with a fully heteroscedastic model. This means that we in some cases we may have proper alternatives that do not involve having a more complex model to handle heteroscedastic error terms. Since post-processing approaches are faster, easier to implement and depend on very few weak assumptions, they constitute a promising area for Big Data applications that warrants further exploration and will be a direction of future work. Furthermore, with QR we can now have prediction intervals for all kinds of machine learning algorithms, even non-probabilistic ones.

We showed that accounting for heteroscedasticity can greatly improve our confidence over traditional pointwise predictions by generating heteroscedasticity-consistent confidence and prediction intervals using different approaches. As it was mentioned in this paper, within the transportation systems context, from the user’s perspective, it is often of extreme importance to have precise estimates which can somehow incorporate the instant volatility that features a given moment. Hence, as a future line of work, we will apply the late presented procedures to real data such like public transport travel-times data.

References


Considerations for Using Big Data in Sustainable Plan Making: Citizen Generated Data for Bike-share Planning

Nader Afzalan

Ph.D. Candidate, College of Architecture and Planning, University of Colorado Denver
Lecturer, Environmental Design Program, University of Colorado Boulder
Vice-chair, American Planning Association- Technology Division
ENVD, Urban Futures Lab, University of Colorado, 314 UCB, Boulder, CO 80309, USA
nader.afzalan@colorado.edu

Abstract

Sustainable planning processes require the participation of multiple actors at multiple scales. Considering the value of web-based technologies in generating diverse and local knowledge, local governments and planning organizations are increasingly using these technologies for citizen engagement. These technologies generate big, detailed, and unstructured data about citizens’ interests and their knowledge of the environment. However, our understanding of usability of this data in plan-making is still limited. Focusing on the City of Cincinnati Bike-share planning, as an example of a sustainable planning effort, this study explores how a web-GIS participatory tool is used for retrieving local knowledge, and how this local knowledge is incorporated in the plan. Building on the literature on local knowledge and participatory planning, this study addresses the usability of crowdsourced Volunteered Geographic Information (VGI) in plan-making processes. Employing spatial and content analysis methods, it analyzes people’s suggestions and examines how those suggestions are incorporated in the plan. Unstructured interviews provided insights on how citizens’ suggestions are incorporated in the final plan and what where the opportunities and limitations of doing such. Variety of factors affected incorporation of the generated big data in Cincinnati bike-share plan: the scope and type
of the plan, the capacity of the planning organization in analyzing data and facilitating peoples’ participation, and the capacity of the online tool in providing valid and relevant information.

**Keywords:** Big data, Sustainable Plan-making, Local Knowledge,Volunteered Geographic Information, Organization Capacity, Bike-share.

**Introduction**

Sustainable planning processes require participation of multiple actors at multiple scales. Considering the value of local knowledge in these processes, local governments and planning consultants are increasingly using web-based technologies to learn from different citizens at multiple scales. These technologies generate big, detailed, and unstructured data about citizens’ interests and their knowledge of the environment. However, our understanding of usability of this information in plan making is still limited.

Focusing on the City of Cincinnati Bike-share planning as a case study, this study explores how a web-GIS participatory tool is used for generating local knowledge and how is this local knowledge incorporated in the plan? Building on the literature on the use of local knowledge in participatory planning, this study addresses the usability of crowdsourced Volunteered Geographic Information in plan-making processes. Employing spatial and content analysis methods, it analyzes people’s suggestions and examines how those suggestions are incorporated in the plan. Unstructured interviews with the project manager, provided insights on how citizens’ suggestions are incorporated in the final plan and what where the limitations of doing such.
Background

Participatory plan-making has been discussed through different lenses, including the importance of dialogue and consensus building (Forester, 1989; Innes & Booher, 2010), the role of local knowledge and its interaction with power in professional planning practice (Corburn, 2003; Fischer, 2000; Forester, 1989; Hoch, 2007), and the importance of incorporating local knowledge in planning processes (Corburn, 2003). Local knowledge is the core value in participatory processes, either as a medium for consensus building among citizens or a medium for decision-making. This study considers local knowledge as a general construct that can be used and incorporated in plan-making and as a driver for decision-making (Yli-pelkonen & Kohl, 2005). However, it does not consider local knowledge as a medium for supporting dialogue or facilitating mutual understanding. Local knowledge can be simply defined as folk culture (Brush & Stabinsky, 1996) that is attached to socio-cultural contexts (Fischer, 2000; Yanow, 2004; Corburn, 2005) and is different from scientific or expert knowledge, which is based on reproducible and abstract procedures (Ezrahi, 1990). The difference between local and scientific knowledge is about the method of data collection, criteria and standards for evidence, and techniques of information analysis (Corburn, 2003). Local or lay knowledge is constantly negotiated among and regenerated by locals. It is not easily quantifiable and measurable. Chapter 26 of Agenda 21 emphasizes the importance of local knowledge and focuses on the “recognition of [local peoples’] values, traditional knowledge and resource management practices with a view to promoting environmentally sound and sustainable development” (Quarrie, 1992). Using local knowledge can lead to a more adaptive process by providing context and experience-based information (Booher & Innes, 2002), supporting interactions at the community level and increasing adaptive capacity (Mercer, Kelman, Taranis, & Suchet-Pearson, 2010), providing cost-effective processes (Healey, McNamara, Elson, & Doak, 1988), or bridging the gap between citizens, experts and community organizations (Rantanen, 2007).

The popularity and accessibility of the Internet has facilitated the production of local knowledge. Web 2.0 has the capacity for producing user generated content (De Longueville, (Bishr & Mantelas, 2008& Coleman, Georgiadou, Labonte, Observation, & Canada, 2009, De
Longueville, 2010) and harnessing the collective intelligence of communities (O’Reilly, 2007). In addition, with the integration of web 2.0 and GIS, various forms of spatial information are now voluntarily generated by users (Flanagin & Metzger, 2008; Hall, Chipeniuk, Feick, Leahy, & Deparday, 2013). While we are dealing with “mass” geo-tagged user-generated data, the application of it in plan-making has not been thoroughly explored (Hall et al., 2013).

The Internet has made information production easier and less expensive, while providing a collaborative environment that allows sharing and combining various experiences and information (Flanagin & Metzger, 2008; Gouveia & Fonseca, 2008). "Web 2.0 is based on three main concepts: user-generated content, the interoperability of information systems and the inclusion of the social context of the user." (De Longueville, 2010). One of the premier components of web 2.0 in spatial platforms is “volunteered geographic information” (VGI) which is introduced by Goodchild (2007) as geospatial content that is being generated by users to meet the needs of various communities. VGI, in addition to other terms such as user generated content (Sieber, 2007) and collaboratively contributed geographic information (Bishr and Mantelas, 2008), emphasizes the participatory nature of GIS. VGI can enhance institutions' decision making by providing qualitative and quantitative local information (Barton, Plume, & Parolin, 2005). This knowledge varies in different contexts since various people with different backgrounds contribute to creating the information (Bishr & Mantelas, 2008) at multiple scales. VGI also contains types of data that have not ever been discovered in traditional mapping. It can be produced either by people who are not formally invited to do so, or those who are invited by a facilitator or through a formal process. Considering the usability of the latter type, Seeger (2008) argues that it can be used by local organizations or governments for sharing spatial information, gathering individuals' ideas that consider existing or proposed situations, and learning about potential sources of tension. It can also add to other methods of data collection, including focus groups, workshops, public meetings and surveys.

There are several constraints with using local knowledge in general and specifically VGI in urban planning processes. It can threaten professionals by allowing people challenge their trustworthiness and legitimacy (Corburn 2005,67). It is also not applicable to all socio-economic
problems at all scales (Fischer, 2000). In addition, user generated information in the online environment has been produced through bottom up approaches and does not rely on top down monitoring processes that control the information quality. It is not filtered; therefore, it may not be very well organized, accurate, or up to date. (Metzger & Flanagin, 2003; Goodchild & Li, 2012; Flanagin & Metzger, 2008; Rieh & Danielson, 2008). Specifically, there are several concerns regarding the quality (Giordano, Liersch, Vurro, & Hirsch, 2010; Hall et al., 2013; Scheuer et al., 2013), credibility (Bishr & Kuhn, 2007; Seeger, 2008; Flanagin & Metzger, 2008), and vagueness of this information (Longueville, B.D., Ostlånder, N., Keskitalo, 2009). Some researchers also argue that using this knowledge may cause issues of privacy, security (Barton et al., 2005), and access to the Internet (Seeger, 2008). Considering all the opportunities and constraints of using VGI, local governments and planning organizations should be prepared to effectively incorporate online planning tools in their current planning processes. Organizational capacity of planning institutions and the capacity of the tool they use for soliciting citizen knowledge affect participatory plan-making processes.

![Diagram](image.png)

**Fig. 1.** Forces that influence the usability of VGI in participatory plan-making

Three elements influence crowdsourcing activates: an organization that benefits from the crowd activity, the crowd, and a platform that hosts the crowd activity and links it to the organization (Zhao & Zhu, 2012). This study focuses mainly on organization and platform
capacity in providing and incorporating crowd-sourced knowledge in plan making, as a process, and plan, as the product. It is not focused on the crowd capacity in providing the knowledge. Online crowdsourcing methods provide opportunities for exploiting crowds’ wisdom (Brabham, 2008, 2009; J Evans-Cowley, 2011; Lévy & Bonomo, 1999) and overcoming some of the issues of the traditional methods of participation, including lack of participants’ diversity (Brabham, 2009) and limitations of time and space (Jennifer Evans-Cowley & Hollander, 2010). However, several considerations should be taken into account for crowdsourcing methods to be successful (Evans-Cowley, 2011): (a) crowdsourcing costs money and time, (b) it is not easy to attract people to participate in the crowdsourcing activity, (c) there is the issue of the digital divide and equality in access to the Internet, (d) both the users and the sponsors require technical support, (e) logging in should be a requirement for users in order to understand who is participating, (f) continuous feedback should provide information for citizens to learn about what is happening, (g) in order to have a successful crowdsourcing activity, the problems should be clear and defined well, (h) crowdsourcing may generate a large amount of responses and data that are not easy to be handled.

The case study

This study focuses on the City of Cincinnati’s effort in using a web-GIS crowdsourcing tool (Shareabouts) for collecting peoples’ ideas about their desired locations for bikeshare stations, mainly in the downtown, Over The Rhine, and uptown area. The bike share program is part of The City of Cincinnati’s goal in providing a new option for mobility around town that is affordable, accessible and visible for citizens and tourists. This tool allowed people to locate points on a map of Cincinnati to suggest locations or to support locations by clicking a support button. In addition, participants were allowed to describe why they proposed a location or participate in a discussion by supporting or opposing others’ ideas (Alta Planning + Design, 2012). The participants were not required to register and provide their personal information such as names or email addresses. During 36 days, 330 locations were suggested by 206 people, a total of 503 comments were made, and 1773 times various locations were supported. 54% of the participants were male, 30% were female and 16% were unknown. The participants’ number
(206) is not low since based on Cincinnati Bicycle Transportation Plan (2010) in 2010 Out of approximately 135,000 employed people over age 16, only 675 were regular bicycle commuters.

**Fig. 2.** The City of Cincinnati’s interactive GIS website interface.

The plan (Cincinnati Bike Share Feasibility Study) includes the following eight sections: introduction, benefits of bike-sharing, expected users, ownerships, local context analysis, system planning, financial assessment, and summary and recommendations. The data collected through crowdsourcing website is used for system planning, primarily for finding citizens’ desired locations for bikeshare stations. Among the five main objectives of this plan, the crowdsourcing website informs the following ones: (a) Evaluate the preparedness of Cincinnati and identify the most suitable areas for bike sharing and any obstacles that could impact success; (b) Identify an initial service area and size for a potential bike share system from which to forecast expected demand, costs and revenues.” (Alta Planning + Design, 2012, 1).

As part of the participatory process, the consultant which was mainly involved in the creation of the plan, arranged several meetings with the City staff, business owners, and similar stakeholders; but, not with citizens. The crowdsourcing website was their main medium for learning about citizens’ ideas and interests regarding the location of bikeshare stations (Interview with Brian, July 2014).
The followings are the main criteria that informed the case study selection. (a) Tool capability: the project uses a webGIS participatory tool that not only provides opportunities for people to express their ideas by locating points on a map but also by creating comments and explaining their intentions and reasons for their suggestions; (b) Plan completion: at the time of the study it was the only plan of its type that was completed and accessible to public; (c) Subject: bike-share planning is an example of new method for incorporating environmental sustainability goals in urban planning process; Data: amount of the generated data is big enough so that the planning organization needs to use analysis methods for making sense of it.

**Methods**

The study employs qualitative and quantitative methods. Spatial and content analysis methods are used in order to identify locations that citizens suggest for having bike-share stations and the reasons they mention for expressing such interest. Qualitative interpretive analysis of the plan allowed the researcher to evaluate in which parts of the plan and how the generated information is incorporated in the plan. In addition, two indepth semi-structured interviews with a planner who was involved in using the tool and creating the plan provided information regarding how and why the crowdsourced knowledge informed the plan creation.

*Content analysis:* To explore why citizen like or dislike having bike-share stations in the suggested locations, this research analyzes participants’ comments by using a content analysis method. Considering peoples’ comments as the unit of analysis, this study implements content analysis to “interpret meaning[s] from the content of text data” (Hsieh & Shannon, 2005, 1277) by exploring all the 503 collected comments.

This study used content analysis software to find and categorize the repeated themes in the participants’ comments. It uses the software to compute the rate at which some particular words or phrases are seen in the comments and in the whole discussion. The software helped with determining themes by finding similar words or phrases and grouping them. For example, it searches for similar words such as *road, trail, path* and puts them on a same category. This analysis provides a basis for detection and categorization of the repeated themes. The software
also helped with qualitatively coding the comments and re-checking the identified themes. Since the users’ informal communication was filled with slang and contextual information, this qualitative step was crucial on top of the automated analysis.

Spatial analysis: To identify the users’ most desired locations and areas for the placement of the bike-share stations, ArcGIS software is used. The software used Kernel Density, a spatial analysis tool that basically computes the density of each feature in a neighborhood around them, for identifying the clusters of citizens suggested locations.

Where and why citizens like to have bike-share stations in Cincinnati?

The two tables below demonstrate why people like to have bike-share stations, by showing the categorized themes, sub-themes and groups.

**Table 1. Results of the content analysis: categorized Themes**

<table>
<thead>
<tr>
<th>Theme</th>
<th>Accessibility</th>
<th>Replace Trips That Would Otherwise Be Made On Foot</th>
<th>Riding up the hill</th>
<th>To be Green</th>
<th>Negative Effects On Businesses</th>
<th>Avoid Being Stuck In Traffic</th>
<th>Avoid Parking Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage</td>
<td>83.50%</td>
<td>6.60%</td>
<td>4.10%</td>
<td>2.10%</td>
<td>1.30%</td>
<td>1.30%</td>
<td>1%</td>
</tr>
</tbody>
</table>

The majority of the participants (83.5%) give “accessibility” to a particular location as the main reason for suggesting a bike-share station location. Each of the themes above are also coded into sub-themes and groups. The table below demonstrates these sub-themes and groups in detail.

**Table 2. Results of the content analysis: categorized Themes, Sub Themes, and Groups**

<table>
<thead>
<tr>
<th>Theme</th>
<th>Sub-Theme</th>
<th>Percentage of each sub-theme</th>
<th>Percentage of each group</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility</td>
<td>Access to Commercial Units</td>
<td>34.10%</td>
<td>8.9%</td>
<td>Access and Proximity to businesses (General)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.5%</td>
<td>Access to a parking area</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5.2%</td>
<td>Access to a restaurant, cafe or a bar</td>
</tr>
<tr>
<td>Category</td>
<td>Percentage</td>
<td>Description</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------------</td>
<td>-----------------------------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Downtown Accessibility</td>
<td>16.20%</td>
<td>Access to Downtown</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access within the Downtown</td>
<td>5.8%</td>
<td>Access to Downtown</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access from Downtown</td>
<td>1.3%</td>
<td>Access from Downtown</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to College or University</td>
<td>8%</td>
<td>Access to College or University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to a Bus Stop or a Metro Station</td>
<td>3.7%</td>
<td>Access to a Bus Stop or a Metro Station</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to Bike Trails and Paths</td>
<td>2%</td>
<td>Access to Bike Trails and Paths</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to Municipality City Hall Court</td>
<td>2%</td>
<td>Access to Municipality City Hall Court</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to Cemetery</td>
<td>.2%</td>
<td>Access to Cemetery</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to Cultural Activities and Sports</td>
<td>3%</td>
<td>Access to Cultural Activities and Sports</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to Fountain Square</td>
<td>2.4%</td>
<td>Access to Fountain Square</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to parks</td>
<td>9.5%</td>
<td>Access to parks</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to Offices</td>
<td>3%</td>
<td>Access to Offices</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neighborhood and Community Access (General)</td>
<td>6.3%</td>
<td>Neighborhood and Community Access (General)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Access to New Port</td>
<td>1.5%</td>
<td>Access to New Port</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replace Trips That Would Otherwise Be Made On Foot</td>
<td>2.6%</td>
<td>Helps people bike instead of walk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>To park and walk from here</td>
<td>1.5%</td>
<td>To park and walk from here</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Riding up the hill</td>
<td>1.3%</td>
<td>Riding up the hill</td>
<td></td>
<td></td>
</tr>
<tr>
<td>To be Green</td>
<td>1.3%</td>
<td>Reduce Car Use</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Being Green</td>
<td>.8%</td>
<td>Being Green</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Negative Effects On Businesses</td>
<td>6.7%</td>
<td>Negative Effects On Businesses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avoid Being Stuck In Traffic</td>
<td>.8%</td>
<td>Avoid Being Stuck In Traffic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avoid Parking Fee</td>
<td>1.3%</td>
<td>Avoid Parking Fee</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Proc. NSF Workshop on Big Data and Urban Informatics, Nov. 2015  Page 303
Table 3. The portion of each group compared to the other ones.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access to Cemetery</td>
<td>0.20%</td>
</tr>
<tr>
<td>Avoid Being Stuck In Traffic</td>
<td>0.80%</td>
</tr>
<tr>
<td>Being Green</td>
<td>0.80%</td>
</tr>
<tr>
<td>Avoid Parking Fee</td>
<td>1.30%</td>
</tr>
<tr>
<td>Reduce the Car Use</td>
<td>1.30%</td>
</tr>
<tr>
<td>Riding up the hill</td>
<td>1.30%</td>
</tr>
<tr>
<td>Access from Downtown</td>
<td>1.30%</td>
</tr>
<tr>
<td>Findlay Market</td>
<td>1.30%</td>
</tr>
<tr>
<td>To park and walk from here</td>
<td>1.50%</td>
</tr>
<tr>
<td>Access to New Port</td>
<td>1.50%</td>
</tr>
<tr>
<td>Access to Municipality, City Hall, or Court</td>
<td>2%</td>
</tr>
<tr>
<td>Access to Bike Trails and Paths</td>
<td>2%</td>
</tr>
<tr>
<td>Access to Hotels and Meeting Centers</td>
<td>2%</td>
</tr>
<tr>
<td>Access to Fountain Square</td>
<td>2.40%</td>
</tr>
<tr>
<td>Helps people bike instead of walk</td>
<td>2.60%</td>
</tr>
<tr>
<td>Access to Offices</td>
<td>3%</td>
</tr>
<tr>
<td>Access to Cultural Activities and Sports</td>
<td>3%</td>
</tr>
<tr>
<td>Replace Trips That Would Otherwise Be Made On Foot</td>
<td>3%</td>
</tr>
<tr>
<td>Access to Hospitals</td>
<td>3.60%</td>
</tr>
<tr>
<td>Access to a Bus Stop or a Metro Station</td>
<td>3.70%</td>
</tr>
<tr>
<td>Access to a parking area</td>
<td>4.50%</td>
</tr>
<tr>
<td>Access to a restaurant, café, or a bar</td>
<td>5.20%</td>
</tr>
<tr>
<td>Access within Downtown</td>
<td>5.80%</td>
</tr>
<tr>
<td>Neighborhood and Community Access</td>
<td>6.30%</td>
</tr>
<tr>
<td>Access to Downtown</td>
<td>6.50%</td>
</tr>
<tr>
<td>Negative Effects On Businesses</td>
<td>6.70%</td>
</tr>
<tr>
<td>Access to College or University</td>
<td>8%</td>
</tr>
<tr>
<td>Access and Proximity to businesses (General)</td>
<td>8.90%</td>
</tr>
<tr>
<td>Access to or within parks</td>
<td>9.50%</td>
</tr>
</tbody>
</table>
As we see in fig. 3, participants’ demands regarding bike-share stations are not high only in the downtown area, but also in the uptown area and some other neighborhoods far from the downtown.
How are the citizens’ ideas incorporated in the plan?

Based on the created plan and interviews with a planner who was involved in plan-making, the website is used primarily for identifying suitable locations for bike-share stations (System Planning section of the plan). The plan’s suggested locations for bike-share stations is strongly correlated with the suggested locations by citizens. Only 5 of out 38 (13%) suggested stations by the plan are located outside of the main clusters of citizens’ suggestion for bikeshare stations (see fig. 4). These 5 suggested stations are all located in Uptown Area. In addition, based on the collected geo-tagged and text-based information (interview with Brian, July 2014) the plan introduces number of places as the highest supported ones, including Washington Park, Fountain Square, Findlay Market. These places also correlate with peoples’ suggestions. However, the plan does not introduce the participants’ priorities or desired types of activities (e.g. access to parks, businesses, university etc.) that support their suggested locations.

The participants’ text-based comments are reviewed qualitatively and used to learn about the reasons that people suggest those locations (interview with Brian, July 2014). However, there is no direct reference to these comments in the plan. Brian introduces the following points as the reasons for not using specific methods of content or sentiment analysis: lack of staff and time, and not having enough reasons for doing so. He clarifies that he is not sure if using those methods is part of their priorities in their plan making process.
Fig. 4. Suggested locations by the plan
Discussion:

The online crowdsourcing tool is primarily used for crowdsourcing citizens’ ideas and introducing the project to public. It provided an opportunity for planners to easily learn about citizens’ desired locations for bike-share stations by visualizing them on a map. Brian believes that citizens’ suggested locations in the downtown and uptown area were not different from the locations that planners would choose even without having citizen participation. Therefore, the generated local knowledge has not been useful to the planners as a new set of “information”. However, it has been useful to the planners for validating their decision-making and making sure that their decision is consistant with citizens’ desire (Interview with Brian, July 2014). On the other hand some of the unexpected information that they received through the site, including peoples’ considerable interest in having bikeshare stations outside of the area defined by the project, were very helpful for the planners to learn about citizens’ interest in expanding the bike-share system and plan for the next stages (Interview with Brian, July 2014).

There are several limitations with the ways in which the crowdsourcing tool is used in this plan. The validity and quality of the peoples’ comments in not examined. For example, 6.7% of the comments belongs to one of the online participants who has expressed her objection regarding more than twenty of the stations suggeted by other participants; however, our interview with the planner indicates that he was not aware of such action (Interview with Brian, July 2014). No one from the planning team was assigned to facilitate peoples’ online interation too. In addition, annonymity of the online participants introduces other considerations in using this information in the plan-making process specially since the online crowdsourcing tool is used as the main medium for soliciting public input. Although planners consider the generated informaiton very valuale due to its quantity and quality (being detailed-oriented), there is still the question of representativeness; to what extent this information represents the whole community?

The plan did not use the the collected text-based informaiton effectively. It did not consider peoples’ suggested priorities in the plan. For example, The content analysis results show that having access to downtown areas from other parts of the town is one of the five most desired suggestions made by the online participants. However, the plan suggests that first the
downtown stations should be placed and then the uptown ones. This study does not argue that the online comments should be fully incorporated in the plan; however, it argues that conducting structured methods for analyzing the generated textual data unveils patterns and hierarchies that planners may find valuable. Not only staff and time limitations affected ways in which planners used the generated information in the plan, but also their knowledge of using the data and their perception of the data usability.

**Conclusion:**

Spatial online participatory tools facilitate the generation of local knowledge from diverse people and at various scales; however, the usefulness of this knowledge depends on variety of factors including the scope and type of the plan, the capacity of planning organizations in analyzing big data and facilitating peoples’ participation, and the tool capacity in providing valid and relevant information. The generated knowledge may be used by planners not as a new information that informs plan-making, but as a dataset that validates this process.

Future studies are required to examine the extent in which various local governments and planning organizations are equipped with the required skills and technologies to analyze big data generated through online tools and social media forums and incorporate the data in their planning processes. In addition, future research is needed to examine the opportunities and constraints that citizens face when engaging in planning processes using online participatory tools.

**References:**


Planning for the Change: Mapping Sea Level Rise and Storm Inundation in Sherman Island Using 3Di Hydrodynamic Model and LiDAR

Yang Ju¹, Wei-Chen Hsu¹, John D. Radke¹, ², William L. Fourt¹, Wei Lang³, Olivier Hoes⁴, Howard Foster⁵, Gregory Biging⁶, Martine Schmidt-Poolman⁵, Rosanna Neuhausler², ⁶, Amna Alruheli¹, William F. Maier⁶

Introduction: In California, one of the greatest concerns of global climate change is sea level rise (SLR) associated with extreme storm events. We argue it is necessary to map storm inundation to understand and plan for future infrastructure changes. A hydrodynamic model, 3Di, by TU-Delft, Netherlands, is used to simulate the inundation of Sherman Island in the Sacramento-San Joaquin River Delta, California. The model simulation is based on 6-minute interval water level data for a 72-hour, near 100-year storm event, coupled with 0.5 meter, 1.0 meter, and 1.41 meter SLR for different inundation scenarios. The big data, high resolution digital elevation model (DEM) from Light Detection and Ranging (LiDAR) is used to describe the ground surface. The output includes

1 Department of Landscape Architecture and Environmental Planning, UC Berkeley, Berkeley, 94720-1820, California, USA;
2 Department of City and Regional Planning, UC Berkeley, Berkeley, 94720-1820, California, USA;
3 Department of Building and Real Estate, the Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong;
4 Faculty of Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628 CN, Delft, Netherlands.
5 The Center for Catastrophic Risk Management, UC Berkeley, Berkeley, 94720, California, USA;
6 Department of Environmental Science, Policy, and Management, UC Berkeley, Berkeley, 94720, California, USA.
a series of inundated areas with 1-hour interval time steps, providing spatial inundation extents and water depth. In addition we combine results from each time step to calculate inundation frequency and average inundation depth for an entire storm event. We demonstrate that when sea level rises more than 1 meter, there are major impacts to Sherman Island. Further we intersect the inundation results with infrastructures on Sherman Island and identify the impact. In all, this work serves as a fine database for better planning, management, and governance to understand future scenarios.

**Keywords:** sea level rise, mapping, 3Di hydrodynamic model, LiDAR, Sherman Island
1 Introduction

In California’s coastal areas, one of the great concerns of global climate change is sea level rise (SLR) associated with extreme high tides (Heberger et al., 2009). By 2100, mean sea level (MSL) will rise between 1.2 m and 1.6 m (Bromirski et al., 2012), and it is assumed this will cause a series of impacts along coastal areas, such as inundation and flooding of coastal land, salt water intrusion, increased erosion, and the decline of coastal wetlands, etc. (Titus et al., 1991; Nicholls and Cazenave, 2010). Among all the impacts, flood risk is likely the most immediate concern for coastal regions.

This threat could be more severe in the Sacramento-San Joaquin Delta, as many of its islands are 3 to 8 m below sea level (Ingebritsen et al., 2000). These islands are protected by more than 1700 km of levees (Mount and Twiss, 2005), with standard cross sections at a height of 0.3 m (1 ft) above the estimated 100-year flood elevation (Ingebritsen et al., 2000). However, with a projected SLR between 1.2 m and 1.6 meters, these current levees could be easily overtopped, and the islands could be flooded.

Several efforts have been undertaken in the San Francisco Bay area to measure and understand the impact of SLR and storm inundation (Knowles et al., 2009; Knowles et al., 2010; Heberger et al., 2009; Bigging et al., 2012). By using GIS software, these studies intersect a water surface with a ground surface to identify inundated areas. The water surface is usually interpolated from measured water level data at existing gauges. The ground surface is usually obtained from LiDAR, providing a high resolution from 1-5 meters. It should be noted that the interpolated water surface is static since it only describes the water surface condition at a particular water level, such as MSL or mean higher high water (MHHW) level. However, real tides and storm events are dynamic processes. The San Francisco Bay area and Sacramento-San Joaquin Delta region are characterized by semi-diurnal tides each day, meaning there are two uneven heights of high tide and low tide, and should be modeled dynamically to simulate all stages in the tidal cycle and the movement of tides during a storm event.
With this in mind our study uses a hydrodynamic model, Delft 3Di, to better simulate the dynamics of tidal interaction during an extreme storm event. In addition, a big dataset, that is a 1 meter resolution digital surface model (DSM), is generated from LiDAR in order to accurately describe the ground surface and to indicate the water pathway for 3Di simulations. We study Sherman Island where significant critical infrastructure exists, and we simulate a near 100-year storm with various scenarios of sea level rise. We analyze and map the inundation extent, frequency, and average depth based on the model outputs.

2 Study Area

Our study area, Sherman Island and its adjacent regions, is located at the confluence of Sacramento River and San Joaquin River (Fig. 1). Sherman Island is one of the major islands in Sacramento–San Joaquin River Delta and is located at the transition from an estuarine system to a freshwater system. Sherman Island has significant infrastructures, including Highway 160, electric high-power transmission lines and natural gas pipelines. According to NOAA, the MSL (1983-2001 epoch) measured at nearby NOAA Port Chicago Gauge is 1.116 m, and MHHW (1983-2001 epoch) is 1.833 m (NOAA Tide and Currents, 2010), based on North American Vertical Datum of 1988 (NAVD 88). The average elevation of Sherman Island is below MSL. Therefore the island is surrounded by extensive levees to protect it from inundation. Even with this levee system, the island still suffers from floods due to levee failures. The most recent levee failure and flooding happened in 1969. The Army Corps of Engineers spent approximately $600,000 in emergency funds to repair, reslope, and regrade the levee break area after the 1969 levee failure. In addition,
the break area still requires constant levee improvements (Hanson, 2009). Even without a levee failure, the island is still at risk in the face of SLR. The lowest point of the levees in the study area is 2.11 m above NAVD 88. If we assume the SLR by 2100 is 1.4 m, then the MSL would rise to 2.25 m and the levees in the study area would be easily overtopped and the entire Sherman Island would be flooded. Considering the importance of Sherman Island’s infrastructure and its vulnerability to SLR, it is a critical region and needs to be studied.

3 Data and Methods

3.1 Overview

To understand the impact of SLR inundation, a water surface and a land surface are required to identify the spatial extent of the inundated areas. We employ a hydrodynamic model, Delft 3Di to simulate a 72-hour, near 100-year storm associated with 0.5, 1.0, 1.41 m SLR scenarios. We create a Digital Surface Model (DSM) generated from airborne LiDAR to capture the elevation of important ground objects, e.g. levees. The model output is a time-series of inundations, providing the spatial extent, inundation depth, and water level. Our workflow is shown in Fig. 2. All elevation data uses the North American Vertical Datum 1988 (NAVD 88).

![Fig. 2 Work flow of this study](image)

3.2 3Di-hydrodynamic model
The 3Di hydrodynamic model (Stelling, 2012), developed by TU-Delft, Netherlands, dynamically simulates the movement of tides through a digital ground surface. The inputs of the model include time-series water level data and ground surface data. The output of the model is a time-series of simulated inundation, providing inundation extent, depth, and water level. The time interval of the outputs is defined by the user. By further processing, the user generates inundation frequency and average inundation depth by combining results from each time step. With an additional dimension of time, this model simulates the dynamics of flood and helps to identify most vulnerable locations from SLR and storm inundation. In addition, by combing the output from each time step, an inundation animation can be created to provide visual communication for the general public; or to improve communication outside of scientific community, making it a great educational tool.

3.3 Water level data

The first step of the 3Di input is water level data. To estimate the impact of a worst case scenario, we choose a near 100-year storm event as the baseline, and then increase levels of SLR increments are added to the baseline. Being a dynamic model, 3Di requires time-series water level data for the entire storm event as input. However, existing 100-year storm calculation methods and studies (NOS CO-OPS, 2013) only provide estimates of water levels for a single stage such as the mean sea level (MSL), mean higher high water (MHHW) level, and mean lower low water (MLLW) level. Considering this fact, a historic storm whose peak water level is close to 100-year storm is used as the water level input. Shown in Tab. 1, two storms that occurred in 1983 exceed the estimated 100-year storm at San Francisco NOAA tide station (NOAA ID: 9414290), and a third highest storm occurred on Feb. 6, 1998, with peak water level close to the estimated 100-year storm (NOS CO-OPS, 2013). All these three extreme storms occurred during El Niño years.
Tab. 1 Estimated and historic extreme storms at San Francisco and Port Chicago NOAA gauge

<table>
<thead>
<tr>
<th>Station Name</th>
<th>Date</th>
<th>Estimated 100-year storm (m)</th>
<th>Peak Water Level (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>San Francisco</td>
<td>01/27/1983</td>
<td>2.64</td>
<td>2.707</td>
</tr>
<tr>
<td></td>
<td>12/03/1983</td>
<td></td>
<td>2.674</td>
</tr>
<tr>
<td></td>
<td>02/06/1998</td>
<td></td>
<td>2.587</td>
</tr>
<tr>
<td>Port Chicago</td>
<td>02/06/1998</td>
<td>-</td>
<td>2.729</td>
</tr>
</tbody>
</table>

Considering the storm’s peak water level and availability of data, the Feb. 6, 1998 storm is selected as the storm to be simulated. More specifically, this study simulates this storm event over 72-hours, from Feb 5 to Feb 7, 1998, in order to allow the model simulation to capture the complete storm dynamic moving through the study area. Furthermore, the water level data used for the 3Di simulation is retrieved from the nearby NOAA Port Chicago gauge (NOAA ID: 9415144), providing measured water level with 6-minute interval in NAVD88.

As for the SLR scenarios, Cayan et al. (2009) and Cloern et al. (2011) studied the projected water level at Golden Gate, and found that the hours of sea levels exceed the 99.99th historical percentile of water elevation would increase to 15,000 per decade by year 2100. And the 99.99th historical percentile is 1.41 m. Thus, this study assumed that 1.41 m would be the SLR increment by year 2100. This study also analyzed scenarios of SLR increment of 0.5 and 1.0 m. The SLR increments were added on top of the baseline water level to simulate each SLR scenarios.

3.4 Ground Surface Data

The second input for the 3Di model is a high spatial resolution DSM. The DSM was constructed based on LiDAR data, which uses optical remote sensing technology and can measure the distance to target by illuminating the target with light pulses from a laser (Wehr and Lohr 1999). The density of the LiDAR data is 1 point per 0.7 m2, and there are approximately 140 million points covering the study area. The DSM obtained from LiDAR in this study was originally 1 m resolution, and is resampled to 4 m resolution to meet the compute processing limitations. Even though the DSM was aggregated to 4 m, this spatial resolution still accurately describes the actual ground surface, showing objects such as levees, ditches, buildings, and the pathways that water moves through.
The 3Di model has limitations in the total number of grids that it can process, and it uses a quad-tree approach to reduce the total number of grids for model computation. The quad-tree is a data structure that is based on the regular decomposition of a square region into quadrants and sub-quadrants (Mark et al., 1989). 3Di draws finer quadrants when elevation changes greatly within a short x,y distance. Considering Sherman Island is relatively flat and the only abrupt change in topography is the levee, only levee data are included into the model to create finer grids, and coarser grids are generated for the rest of the homogeneous study area. The DSM and the quad-tree grids for Sherman Island are shown in Fig. 3.

![DSM and quad-trees](image)

Fig. 3 DSM and quad-trees, showing 3Di draws finer grids along the levees and coarser grids for the rest areas.

4 Results

The 3Di simulation output is a time-series of inundated areas with an output time interval that is defined by the user. Each output provides the spatial extent of inundation and depth. In this study, the time interval is set as 1 hour, and a total of 72 outputs are generated from the model. Based on the time-series outputs, this study also analyzes inundation frequency and average inundation depth. Fig. 4 is an example showing the inundation extent and depth in hour 1, 24, 48, 72 for the simulated near 100-year storm associated with 1.41 m SLR.
100-year storm+1.41m SLR

Fig. 4 An example showing simulated inundation from a 72-hour, 100-year storm associated with 1.41 m SLR, showing inundation extent and depth in hour 1, 24, 48, 72, respectively.

4.1 Inundation extent

The results show that during a 100-year storm, a total of 14.68 km$^2$ of land is inundated in the study area. With 0.5m SLR, a total 20.67 km$^2$ of land is inundated, with 1.0m SLR, a total of 57.87 km$^2$ of land is inundated, and with 1.41 m SLR, a total of 72.43 km$^2$ of land is inundated, and summarized in Tab. 2. The inundation extent for different SLR scenarios is shown in Fig. 5. Here the western end of Sherman Island is constantly underwater in all the modeled SLR scenarios as it is not protected by levees. In 0.5 m SLR scenario, only minor inundation occurred in the rest of the island. In 1.0 m SLR scenario, over half of the remaining Sherman Island is inundated. In the 1.41 m SLR scenario, the entire Sherman Island is inundated, showing in red color. The simulation shows that when sea level rises above 1.0 m, it will cause major flood impacts on Sherman Island.

Tab. 2 Statistical summuray of inundation by a 100-year storm with different level of SLR

<table>
<thead>
<tr>
<th>SLR (m)</th>
<th>Inundated Area (km$^2$)</th>
<th>Area by inundation frequency (km$^2$)</th>
<th>Area by average inundation depth (km$^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Low (0-0.21)</td>
<td>Medium (0.22-0.64)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low (0-1.98 m)</td>
<td>Medium (1.99-4.01 m)</td>
</tr>
<tr>
<td>0</td>
<td>14.68</td>
<td>4.40</td>
<td>4.99</td>
</tr>
<tr>
<td>0.5</td>
<td>20.67</td>
<td>3.26</td>
<td>7.33</td>
</tr>
<tr>
<td>1</td>
<td>57.87</td>
<td>6.82</td>
<td>23.93</td>
</tr>
<tr>
<td>1.41</td>
<td>72.43</td>
<td>2.37</td>
<td>15.52</td>
</tr>
</tbody>
</table>
4.2 Inundation frequency

Storm is a dynamic process, and impacted areas may not permanently be under water during the entire storm event. Thus, this study then analyzed inundation frequency by using (1) and (2):

\[
I_{x,y,i} = \begin{cases} 
1, & \text{inundated} \\
0, & \text{not inundated} 
\end{cases}
\]  

(1)

\[
F_{x,y} = \frac{\sum_{i=1}^{n} I_{x,y,i}}{n}
\]

(2)

where \(I_{x,y,i}\) is whether raster cell in column \(x\), row \(y\) gets inundated at hour \(i\), \(F_{x,y}\) is the inundation frequency for raster cell in column \(x\), row \(y\), and \(n\) is the total number of outputs, which equals to 72 in this study since a 72-hour event was simulated.

The inundation frequency calculated here is the percentage of hours each piece of land (4 m × 4 m) gets inundated in the entire 72-hour storm event. This study then classifies the inundation frequency in 1.41 m SLR scenario using a natural breaks method. From this classification, low frequency is 0-0.21, medium frequency is 0.22-0.64, and high frequency is 0.65-1. The results from other scenarios are classified using the 1.41 m SLR scenario classification in order to compare across different results consistently. The inundation frequency is shown in Fig. 6, Fig. 7, Fig. 8, Fig. 9 for 0, 0.5, 1.0, 1.41 m SLR, respectively, and a statistical summary is shown in Tab. 2. From the results, it is known that when sea level rises, low frequency areas decrease while high frequency
areas increase, showing that more land would be permanently inundated in the future when sea level rises.

Fig. 6 Inundation frequency during the 3-day 100-year storm associated with 0 m SLR (left)

Fig. 7 Inundation frequency depth during the 3-day 100-year storm associated with 0.5 m SLR (right)

Fig. 8 Inundation frequency during the 3-day 100-year storm associated with 1.0 m SLR (left)

Fig. 9 Inundation frequency during the 3-day 100-year storm associated with 1.41 m SLR (right)

4.3 Average inundation depth
This study also analyzes the average inundation depth, as the inundation depth on top of each piece of land varies in a storm event. The average inundation depth is calculated by (3):

\[
D_{x,y} = \frac{\sum_{i=1}^{n} d_{x,y,i}}{n}
\]

where \(D_{x,y}\) is the average inundation depth (m) at raster cell in column \(x\), row \(y\), \(d_{x,y,i}\) is the inundation depth at raster cell in column \(x\), row \(y\), at hour \(i\), and \(n\) is the total number of outputs, which equals 72 in this study.

Similarly, we classify the average inundation depth in 1.41 m SLR scenario by using the natural breaks method. From this classification, low depth is 0-1.98 m, medium depth is 1.99-4.01 m, and high depth is 4.02-13.22 m. The results from other scenarios are classified using the 1.41 m SLR scenario classification, in order to compare across different results. The average inundation depth is shown in Fig. 10, Fig. 11, Fig. 12, Fig. 13 for 0, 0.5, 1.0, 1.41 m SLR, respectively, and a statistical summary is shown in Tab. 2. The results show that in 0.5 and 1.0 m SLR scenarios, the majority of inundated areas are under low inundation depth, and when it comes to 1.41 m SLR, more lands are under medium and even high inundation depth.

![Fig. 10 Average inundation depth during the 3-day 100-year storm associated with 0 m SLR](image1)

![Fig. 11 Average inundation depth during the 3-day 100-year storm associated with 0.5 m SLR](image2)
5 Discussion and conclusions

5.1 Implication for planning

This study creates a SLR and storm inundation dataset for Sherman Island and its adjacent areas. This is an important and initial step for policy makers, planners, and the public to understand the magnitude and spatial distribution of SLR and storm inundation. The results show that with more than 0.5 m SLR, the levees protecting Sherman Island are overtopped. With 1.0 m SLR, nearly half of Sherman Island is fully inundated, and with 1.41 m SLR, the entire island is inundated. Based on this study, SLR impacts are very significant, especially when SLR is greater than 1.0 m. Local governments can use inundation water level information from the dataset created to improve the levee system and accommodate the rising sea level by constructing new levees to protect areas with high inundation frequency. This dataset can be employed in a suitability analysis for Sherman Island to identify areas with higher inundation risks, and to improve infrastructure planning and/or adopt different planning strategies for the rising sea level.
Our research group is currently studying the SLR impacts on infrastructures, such as pipelines and roads. These infrastructures are designed to allow certain level of inundation, but this tolerance is limited. To better understand SLR and storm inundation impacts on these critical infrastructures, it is beneficial to know the duration and the depth of water sitting on top of any infrastructure. Static models have limitations as they only depict one stage of inundation, where the information on duration and flood dynamic is lost. The dynamic model implemented here provides the additional dimension of time. Subsequent studies from our research group are intersecting the inundation dataset with infrastructure datasets to calculate the duration the infrastructure is impacted, as well as the amount of water sitting on top of it. Further research will identify when the infrastructure gets impacted, estimate the cost, and provide more detailed planning suggestions.

5.2 Limitation and application

This study has several limitations. First, the Sacramento-San Joaquin Delta region has a complex hydrologic system which is influenced by both the ocean and the rivers, making it difficult to conduct hydrologic modeling. Considering that Sherman Island is close to the mouth of the Delta, this study simplified the actual process and assumed that the island is only affected by a tidal surge from the ocean. With the discharge from the Sacramento and San Joaquin River, the simulated process could be different. Second, the model does not incorporate other factors, such as subsidence, sediment deposition, wind, and rainfall, etc. Being an “artificial” system, the Sacramento-San Joaquin Delta region has limited sedimentation and significant subsidence issues that would further exacerbate the impact of SLR inundation. As a result, the 3Di model might underestimate the SLR impacts on Sherman Island. Third, the water level data used in this study could be inaccurate, as there is no gauge currently available in the immediate region of the study area. Finally, the 3Di model has computing limitations that limit the number of grids processed to approximately 125,000 for each simulation. Our study lowered the DSM resolution, from original 1 m to 4 m, to accommodate the computing limitations. As a result, some topographic information, such as smaller ditches and roads, may not be reflected in the model.
No GIS model perfectly represents reality (Fazal, 2008), and inundation models are usually a simple but effective method that identifies inundated areas (Tian et al., 2010). Inundation models provide the possibility to incorporate different datasets and generate models for planners, policy makers and the public to clearly see potential impacts. Compared to previous studies, our study provides a more detailed level of information, and serves as a basis for future analysis. Our research group continues to generate similar datasets for the entire Sacramento-San Joaquin Delta and the San Francisco Bay Area.

Reference


6. Hanson, J. C. 2009. Reclamation District 341, Sherman Island five year plan, SH 08-3.0. Available at: http://ccrm.berkeley.edu/resin/pdfs_and_other_docs/background-lit/hanson_5yr-plan.pdf


14. NOAA Tide and Currents. 2010. Datums for 9415144, Port Chicago CA. Available at: http://tidesandcurrents.noaa.gov/datums.html?units=1&epoch=0&id=9415144&name=Port+Chicago&state=CA


Parameterizing Land Use Planning  
Deploying quantitative analysis methods in the practice of city planning  

Talia Kaufmann, Master in City Planning Candidate 2014  
Department of Urban Studies and Planning, MIT  

The practice of land use planning determines the quantities and locations of various land uses we find in a city. This planning process first sets population thresholds for all city amenities and then turns to distribute them spatially across the city. Think of the practice of land use planning as being similar to playing a Lego game, where players choose a set of building blocks and then consider the multiple options to assemble them together. However, unlike a regular Lego game, in land use planning the players’ role is to determine the rules and allotments that will limit the assembly options of the Lego pieces in a way that will benefit as many interests as possible.

In the Lego game analogy, when players assemble pieces together, they usually follow a physical or mental image of the object they would like to produce. Similarly, when planners face the task of planning a new area or revitalizing an existing one, they explore different urban models and choose the one that best serves the planning purpose. However, the available assessment tools for planners today are mostly qualitative and limited in their ability to allow for comparison. Quantitative analysis methods can equip planners with metrics to assess various urban models and compare their spatial patterns by the specific parameters that characterize them. This research will explore how planners can harness the potential of quantitative urban metrics to develop guidelines and provide recommendations about the quantities and spatial distribution of land uses in cities.

The question of how to determine the spatial organization of lands uses in cities has long been the interest of planners and economic geographers. The idea of modeling the urban environment, whether qualitatively or quantitatively, with the goal of finding the ideal location for land uses given a set of constraints appeared as the solution for optimizing the performance of the built environment in multiple aspects. In recent years, several studies have attempted to establish a predictive quantitative theory of urban organization and development using the vast amount of accessible data while applying methods from complexity science. One such study by Bettencourt et al (2007) presents empirical evidence showing that important demographic, socio-economic, and behavioral urban indicators are, on average, scaling function of a city’s population size that are quantitatively
consistent across different nations and times. Thus, the quantity of each land use type in a city is relative to population size and can be determined by the scaling exponent that is unique for each type. These scaling relations hold no information about the relative position of different uses and their spatial distribution across the city. Nevertheless, land uses often cluster, co-locating to form patterns that may contain information about the uses’ dependency on each other, shared demand, or similar infrastructure and transportation requirements. Therefore, the planning practice is in need of quantitative urban metrics to enable planners with the ability to understand and compare different spatial patterns by the parameters that control them.

This research will analyze a dataset summarizing the fine-grained location of commercial and public land uses in the 50 largest metropolitan areas in the U.S. This dataset will be used to demonstrate how quantitative relationships found in cities can be utilized to produce land use planning guidelines in two main aspects: First, a quantifiable method to assess population levels for different land use types will be demonstrated by implementing the scaling relationships in cities from the Bettencourt et al research (2007). Next, an empirical method for assessing the spatial organization of different cities will be produced by two analysis methods: a clustering analysis to typify different city centers by diameter size and land use composition and a calculation of pairwise distances between uses within city centers. The combination of these analyses will produce a ‘spatial identity card’ for each city including the quantities, type and composition of their city centers as well as a co-location heat map of uses. This identity card will allow for an in-depth assessment and comparison of different urban models to use as metrics and guide the discourse of land use planning.

The Lego game serves as more than an analogy in this case; it is a methodology for the process of planning land uses in cities. A Lego game is an open-ended system that does not try to define a finite product, but rather provides the building blocks and game rules to enable endless possible outcomes. The Lego system can introduce modularity and flexibility to the planning process at the most fine-grain of scales for assembly options. Here, quantitative methods from complexity science to observe universal behaviors across cities can help planners untangle the connections between different land use types from which planning guidelines for assembly options can be constructed. Combining the open-ended manner of a game with the complexity approach of modeling interactions introduces an opportunity to transform the way we practice land use planning.
Estimating Driving and Parking Patterns from Cellular Network Data: An Application to Plug-in Electric Vehicle Charging Infrastructure Planning

**Jing Dong** (Corresponding author)
jingdong@iastate.edu
Department of Civil, Construction and Environmental Engineering
Iowa State University
350 Town Engineering Building
Ames, Iowa 50011
Phone: (515) 294-3957

**Xuesong Zhou**
xzhou74@asu.edu
School of Sustainable Engineering and the Built Environment
Arizona State University
Tempe, AZ 85287, USA
Phone: (480) 9655-827

**Pengfei “Taylor” Li**
pflee2002@gmail.com
School of Sustainable Engineering and the Built Environment
Arizona State University
Tempe, AZ 85287, USA
Abstract

Promoting plug-in electric vehicles (PEV) is considered one of the pathways to sustainable petroleum displacement in the transportation sector. Unlike refueling a conventional diesel or gasoline tank, charging the battery takes a much longer time, from 30 minutes to several hours, depending on the charger power, battery size and its state of charge. Therefore, it is preferred to charge a PEV at the activity destination where the vehicle is parked for a considerable period of time. Understanding the activity-travel patterns of the traveling public becomes a key to design an efficient PEV charging infrastructure system. In general, activity travel information can be obtained from traditional travel surveys, which is a very costly and time-consuming process. Currently, a large volume of smart phone and cell tower based location data are available from emerging Big Data applications. Given a sequence of trajectory data points with time stamps, this paper proposes a new activity pattern identification method that can find the most likely activity types and activity duration at intermediate stops. The proposed algorithm aim to identify home activity duration, major office/work activity duration, rough estimate of departure time and arrival time of trips based on multiday cell phone records from the same user. The resulting travel activity patterns are used to quantify the benefit of offering public charging opportunities and to determine where to site charging stations subject to vehicle travel range constraints. This application can assist policy makers efficiently allocate public resources in aiding the deployment of charging infrastructure.

Keywords

Plug-in electric vehicle (PEV); Activity travel pattern; Space-time network; Hidden Markov model (HMM); Viterbi algorithm.
1. Introduction

Promoting plug-in electric vehicles (PEV) is considered one of the pathways to sustainable petroleum displacement in the transportation sector. The fear that the vehicle has insufficient range to reach the destination, referred to as range anxiety, not only discourages consumer acceptance but also restrains the social benefits of PEVs. This calls for a well-planned charging infrastructure system. Unlike refueling a conventional diesel or gasoline tank, charging the battery takes a much longer time, from 30 minutes to several hours, depending on the charger power, battery size and its state of charge. Therefore, it is preferred to charge a PEV at the activity destination where the vehicle is parked for a considerable period of time. Understanding the activity-travel patterns of the traveling public becomes a key to design an efficient PEV charging infrastructure system. In general, activity travel information can be obtained from direct interview surveys. Household, destination and roadside interview surveys, typically used in the transportation planning analysis, provide valuable samples about the detailed travel activities of each tripmaker, such as the activity location and duration, the mode used and the travel time. Populating activity travel patterns from survey samples, however, is a very costly and time-consuming process.

Currently, a large volume of smart phone and cell tower based location data are available from emerging Big Data applications. On the other hand, typically available vehicle or cell phone location data are still associated with location errors, within a range of 5 meters to 30 meters, or the location data have been randomized or scrambled to protect users’ privacy. A critical data processing challenge in emerging Big Data applications is how to use location data with large errors or only partial information to estimate the most likely activity patterns on the transportation network. Given a sequence of trajectory data points with time stamps, this paper proposes a new activity pattern identification method that can find the most likely activity types and activity duration at intermediate stops. The proposed algorithm aim to identify home activity duration, major office/work activity duration, rough estimate of departure time and arrival time of trips based on multiday cell phone records from the same user.

In this paper, Hidden Markov Model (HMM) is used for travel activity pattern mining in the traffic networks. HMMs are commonly used for speech, handwriting, and gesture recognition, and have been adopted in traffic prediction. For example, Kwon and Murphy (2000) used coupled HMMs to model freeway traffic and predict traffic speed. Their study defined two states
(congestion and free flow) using the mean speed. Statistical inference algorithms were required to train the model, which was not computationally effective even with a small dataset. Qi and Ishak (2014) developed and calibrated HMM using a large amount of real-world traffic surveillance data. The short-term traffic condition was defined in terms of both first- and second-order statistics of traffic speed observations to account for the range, local variation, and trend of traffic over a short time period. Determining destinations from a set of location vectors is a clustering task. There are many options for clustering points, such as k-means, hierarchical clustering techniques, and agglomerative clustering (Krumm and Horvitz, 2006). Coarse-grained location and proximity information from cell phones have been used for activity discovery. For example, focusing on exploring social networks and organizational rhythms of groups, Eagle et al. (2006) used the cell phone data to detect daily and weekly patterns of location transitions. Liao et al. (2007) used GPS sensor data to construct activity models, such as work, leisure, visit, and to identify significant destinations, including home, work, and stores.

2. Space-time network

This section presents a space-time network modeling approach to consider not only the geometry and topology of the road network, but also the time attributes in available location samples. We use a space-time network-based representation that has been widely used in formulating and solving the most likely paths for many dynamic network flow optimization applications, as well as related fields such as measuring transportation space-time accessibility.

Widely used in both transportation geography and transportation network modeling literature, the concept of space-time networks (STNs) aims to integrate physical transportation networks with travelers’ time-dependent movements/trajectories. To construct a STN, the analysis time horizon is first divided into a series of intervals with the same time length $\sigma$. In Figure 1, a physical transportation network is shown on the lower portion, while the upper part plots a set of vertexes, travel arcs and waiting arcs, each with different spatial and temporal characteristics. Specifically, a person, traveling in this illustrative three-node network, departs from origin node $o$ at time $t_0$ and arrives at node $d_1$ at time $t_0 + 2\sigma$. The traveler finally reaches at the third node $d_2$ at time $t_0 + 6\sigma$. The route choice decision can be described as a sequence of arcs traveling at this space-time network, namely the space-time travel arcs $(o, d_1, t_0, t_0 + 2\sigma)$, $(d_1, d_1, t_0 + 2\sigma, t_0 + 3\sigma)$ and $(d_1, d_2, t_0 + 3\sigma, t_0 + 6\sigma)$. The space-time network construct in
Figure 1 (typically used in the literature of transportation network modeling) corresponds to the illustration in Figure 2, where these arcs tracing the movement of an individual form a *space-time path* (Hägerstrand 1970).

![Figure 1 Illustration of physical network and space-time network](image)

![Figure 2 Illustration of a space-time path (Adapted from Hägerstrand 1970)](image)
3. Activity Travel Pattern Estimation

One of important challenges for the proposed trajectory-based activity pattern representation is that there are a large number of possible traffic states for each individual vehicle at any given time, and the partial location data could lead to missing activities, unidentifiable mode, large location matching errors. To aggregate microscopic traffic states (e.g. traveler’s activity and departure time) to macroscopic traffic states (transitions between different activities), we use a Hidden Markov Model (HMM) that uses both geometric data and the topology data in the traffic networks for travel activity pattern mining.

3.1. Hidden Markov Model

Using 30-minute time intervals, a 24-hour day is divided into 48 time intervals. The activity locations are categorized as home (H), work (W) and other (O). After inferring travel destinations from cell phone status and activities, we assign location labels of home (H), work (W), or other (O) to the towers (or the hashed tower IDs). As an example, Figure 3 illustrates a single user’s travel trajectories over 15 days.

Figure 3 Travel activities of an example traveler

HMM assumes that at the beginning of a time interval, the person is at one destination. During the time interval, the person makes exactly one transition between destinations. A self-transition, that is, a transition from a destination to itself, is allowed. The start probability matrix, \( \Pi = \{\pi(d_i, \tau_k)\} \), represents the probability that the individual is at destination \( d_i \) at the beginning of time interval \( \tau_k \). As the individual needs to be at one of the destinations at the beginning of any time interval \( \tau_k \), we have \( \sum_{i=1}^{n} \pi(d_i, \tau_k) = 1 \), for all \( \tau_k \)'s. The transition probability matrix, \( A = \{a(d_i, d_j, \tau_k)\} \), represents the probability that the individual travels from destination \( d_i \) to \( d_j \) during time interval \( \tau_k \). For each \( d_j \) and \( \tau_k \), we have \( \sum_{i=1}^{n} a(d_i, d_j, \tau_k) = 1 \).

Trip chaining data from 2009 National Household Travel Survey (NHTS) (source: U.S. Department of Transportation, Federal Highway Administration, 2009 NHTS, available at http://nhts.ornl.gov) is used compute start probability and transition probability matrices. In the
chained trip file, trips are linked together between two anchored destinations (home, work, and other). The trip chaining dataset provides insight into travel demand based on location, trip purpose, and transportation mode, and can be used to estimate the time and distance related to commuting and other anchored tours.

3.2. Viterbi algorithm

Based on the observations and probability matrix, Viterbi algorithm is employed to find the trip chain (Viterbi path) for the sampled cell phone user. In order to match trajectory observation, the underlying activity-matching algorithm utilizes the Viterbi algorithm to estimate correct activity travel segments as hidden states in HMM. As a dynamic programming solution approach, the Viterbi algorithm is implemented by applying the standard shortest path algorithm to find the most likely path in a Bayesian network. Thus, we can decompose the corresponding shortest path algorithms to different virtual servers to improve the processing speed for a large volume of streaming location data. As a result, a large number of location traces can be mapped and clustered as a space-time trajectory during a typical travel schedule. In addition, the trip chaining data from 2009 National Household Travel Survey (NHTS) dataset are used to construct the Bayesian network.

4. Strategic planning for PEV Charging Infrastructure

After understanding drivers’ activity travel pattern, one potential application is to determine how many charging stations are needed for a given region. An adequate charging infrastructure is considered a technological option for reducing the market barriers to PEVs.

4.1. Analysis Framework

The objectives of strategic infrastructure planning for charging station network include minimizing the number of missed trips for battery electric vehicle (BEV) drivers due to the limited vehicle range, and maximizing the electric miles or energy cost savings for plug-in hybrid (PHEV) drivers. Figure 4 shows the analysis framework for PEV charging infrastructure planning. The dwell times at home, work, and other places, and the travel times (H2W, W2O etc.) calculated based on inferred trip chains are used to evaluate the potential of switching to a PEV and aid the design of the charging infrastructure.
4.2. Charging infrastructure planning for plug-in hybrid electric vehicles

A PHEV’s electricity and gasoline consumptions depend on whether it is operated in CD or CS mode. Under a switching strategy that aims at minimizing the net consumption of gasoline, the vehicle operates in CD mode when the battery’s state of charge (SOC), an indicator of the amount of usable battery energy, is higher than a certain level and operates in CS mode when lower. In this paper it is assumed that a PHEV will operate in CD mode until the CD range is exhausted, at which time the vehicle operates in CS mode. The gasoline and electricity consumption rates when operating in CD and CS modes, in terms of gallon per mile and kilowatt-hours per mile, respectively, are listed in Table 1. Note that PHEVs equipped with a smaller (lighter) battery pack tend to have better fuel economy.

**TABLE 1 PHEV Energy Consumption Rates**

<table>
<thead>
<tr>
<th>CD range (mile)</th>
<th>CD mode</th>
<th>CS mode</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gasoline, $r_{cd,g}$ (gallon/mile)</td>
<td>Electricity, $r_{cd,e}$ (kWh/mile)</td>
</tr>
<tr>
<td>10</td>
<td>0.0095</td>
<td>0.1800</td>
</tr>
<tr>
<td>20</td>
<td>0.0082</td>
<td>0.1975</td>
</tr>
<tr>
<td>40</td>
<td>0.0057</td>
<td>0.2386</td>
</tr>
</tbody>
</table>

The electricity (Equation [1]) and gasoline (Equation [2]) consumptions, as well as the corresponding energy cost (Equation [3]), depend on the energy consumption rates and the
distances traveled in CD and CS mode, which relate to the recharging behavior discussed in the next section.

\[
E_e = \sum_n d_{n,cd} \cdot r_{cd,e}
\]  \hspace{1cm} \text{(1)}

\[
E_g = \sum_n (d_{n,cg} \cdot r_{cs,g} + d_{n,cd} \cdot r_{cd,g})
\]  \hspace{1cm} \text{(2)}

\[
E_c = E_e \cdot c_e + E_g \cdot c_g
\]  \hspace{1cm} \text{(3)}

Whether to recharge at stop \( n \) is determined by charger availability and user decision.

\[
I_n = X_n \cdot Y_n
\]  \hspace{1cm} \text{(4)}

Charger network coverage is represented by the probability that a charger is available at a certain stop. In particular, charger availability, \( X_n \), is drawn from a Bernoulli distribution. \( X_n = 1 \) means a charger is available at stop \( n \); otherwise, \( X_n = 0 \).

\[
\Pr(X = 1) = 1 - \Pr(X = 0) = p, \ 0 \leq p \leq 1
\]  \hspace{1cm} \text{(5)}

When a recharge opportunity is presented, a PHEV driver would consider the recharge urgency, cost, and hassle based on remaining battery capacity, travel schedule, knowledge about the given charger, the next recharge opportunity, and so forth. In the present paper, the user’s decision on recharging is based on the perceived benefit and cost associated with the charging activity, as well as personal preferences. Such a decision-making process is assumed to follow bounded rationality (Simon, 1955); that is, when the benefit-cost ratio is greater than a bounded rationality threshold \( \alpha \), the driver will recharge the battery.

\[
Y_n = \begin{cases} 
1, & \frac{B_n}{C_n} > \alpha \\
0, & \text{otherwise} 
\end{cases}
\]  \hspace{1cm} \text{(6)}

To quantify the benefit and cost associated with a recharge event, the potential energy increase in the battery needs to be determined, constrained by both the remaining battery capacity and the power from the charger.

\[
R_{p,n} = \min(R_{cd} - R_{soc,n}, \frac{P \cdot t_n}{r_{cd,e}})
\]  \hspace{1cm} \text{(7)}

In Equation (7), the vehicle’s full CD range \( R_{cd} \) and electricity consumption rate \( r_{cd,e} \), as well as charger power \( P \), are predetermined parameters. The dwell time \( t_n \) is obtained from the
GPS travel data. The remaining CD range when arriving at stop \( n \) \( (R_{soc,n}) \) can be calculated on the basis of battery level at the previous stop, possible recharge, and travel distance.

\[
R_{soc,n} = \max (0, R_{soc,n-1} + I_{n-1} \cdot R_{p,n-1} - d_n) \tag{8}
\]

The benefit resulting from a particular recharge activity includes gasoline cost savings, refueling burden relief, and possibly other factors, such as driver-perceived environmental benefit. The refueling burden, expressed as a per-gallon equivalent, considers the frequency of refueling and the associated travel time and hassle cost.

\[
B_n = R_{p,n} \cdot r_{cs,g} \cdot (c_g + b_g) \tag{9}
\]

On the other hand, the cost associated with a particular recharge event is composed of electricity cost and the perceived recharge burden, which is quantified by a monetary cost per charge.

\[
C_n = R_{p,n} \cdot r_{cd,e} \cdot c_e + b_e \tag{10}
\]

Note that the benefit and cost are estimated on the basis of current charging potential without consideration of upcoming trips and stops. If future travel activities and charging options are known, drivers might perceive a different benefit or cost, and thus make a different recharging decision. For example, a driver might skip the recharge opportunity when he/she is close to home or anticipates a longer dwell time or a greater charging potential at the next stop.

Based on the quantified benefit and cost, as well as the decision criterion defined in Equation [6], the recharging decision at stop \( n \) can be inferred according to Equation [4]. Finally, the travel distances in CD and CS mode can be computed, as follows, and then plugged into Equations [1]–[3] to estimate energy consumptions and cost.

\[
d_{n,cd} = \min (R_{soc,n-1} + I_{n-1} \cdot R_{p,n-1}, d_n) \tag{11}
\]

\[
d_{n,cs} = d_n - d_{n,cd} \tag{12}
\]

### 4.3. Charging infrastructure planning for battery electric vehicles

Consider a set of candidate sites \( I = \{1,2,\cdots,m\} \) for installing charging stations, and a set of BEV drivers \( J = \{1,2,\cdots,n\} \). The public charger placement problem is to determine the locations and the types of the chargers to be installed in the planning network so as to minimize the number of missed trips, subject to a budget constraint.
Drivers’ travel activities, including trip distances and the dwell time between two consecutive trips, and BEV characteristics, including the electric range and electricity consumption rate, are known. These input variables are defined as follows.

- \( s_{jd(k)} \): Travel distance of driver \( j \)'s \( k \)-th trip on day \( d \) [mile]
- \( t_{jd(k)} \): Dwell time after driver \( j \)'s \( k \)-th trip on day \( d \) [hour]
- \( l_{jd(k)} \): Destination of driver \( j \)'s \( k \)-th trip on day \( d \) [-]
- \( R_j \): Electric range of driver \( j \)'s BEV [mile]
- \( r_j \): Electricity consumption rate of driver \( j \)'s BEV [kW h/mile]

In the case study, the entire fleet is assumed to be BEVs with a 100 mile range (i.e. \( R_j = 100, \forall j \)). And the electricity consumption rate is 300 W h per mile (i.e. \( r_j = 0.3, \forall j \)).

Whether to install an electric vehicle charger at a candidate site or not is denoted as the decision variable.

- \( x_i \): Charger placement at candidate site \( i \in I \) (= 0, if no charger installed; = 1, 2, or 3, if level 1, 2 or 3 charge is installed)

Accordingly, the charging power and cost of each candidate site can be determined based on Table 1. These derived variables are defined as follows.

- \( P_i \): Charging power at candidate site \( i \in I \) is a function of \( x_i \).
- \( C_i \): Charger cost at candidate site \( i \in I \) is a function of \( x_i \).

If a BEV driver’s activity destination is in the candidate sites, that is, \( l_{jd(k)} \in I \), the available charging power is \( P_{l_{jd(k)}} \). If the destination does not belong to the candidate charging station site, or no charger is installed at the candidate site, \( P_{l_{jd(k)}} = 0 \).

When the BEV range is sufficient to finish the driver’s all-day travel activities, that is, \( \sum_k s_{jd(k)} \leq R_j \), We assume that the driver will not use public chargers and only charge the battery when returning home. This assumption is made to simplify the calculation and represents the majority of current BEV adopters’ behavior. It can be relaxed and will not affect the solution. When daily VMT exceeds the BEV range, drivers can take advantage of public chargers and charge the battery at some trip destinations. The energy increase in the battery, measured in miles, can be determined based on the battery’s state of charge, charging power and dwell time at the destination.
Energy increase of the battery from the recharge at the destination of driver $j$’s $k$-th trip on day $d$:

$$R_{jd(k)} = \min \left\{ R_j - R_{soc, jd(k)}, \frac{P_{ld(k)} \cdot t_{ljd(k)}}{r_j} \right\}$$

battery’s pre-charging SOC at the destination of driver $j$’s $k$-th trip on day $d$, which is measured after finishing trip $k$ and before a possible recharging at the destination.

The pre-charging SOC of the BEV at the destination of the $k$-th trip ($R_{soc, jd(k)}$) can be calculated on the basis of battery level at the previous stop, possible recharge, and trip distance.

$$R_{soc, jd(k)} = R_{soc, jd(k-1)} + R_{jd(k-1)} - s_{jd(k)}$$

A negative pre-charging SOC ($R_{soc, jd(k)}$) indicates that the range of the BEV is insufficient to complete the daily travel. Thus, the $k$-th trip and all the subsequent trips on the travel day are considered as missed trips. Let $y_{jd}$ denote the number of missed trips for driver $j$ on day $d$. Thus, the objective function can be written as minimizing the total number of the missed trips of all the BEV drivers on all the travel days.

$$\min f(x) = \sum_{j} \sum_{d} y_{jd}$$

The total cost of building the charging infrastructure needs to be within the maximum allowable budget. The budget constraint is written as follows.

$$\sum_{i} C_i \leq B$$

$B$ The total budget for installing chargers in the entire study area

5. Concluding Remarks

The advances in sensing and communications technologies allow for tracking individual vehicle or person and collect fine-grained spatial and temporal data for travel survey purposes. Recently, a large volume of smart phone and cell tower based location data are also available from emerging Big Data applications. By leveraging spatio-temporal travel survey data and cellular network data, we proposed a methodology to identify activity travel patterns and developed realistic behavioral models that describe travelers’ driving and parking activities. The proposed activity pattern identification method is used to quantify the benefit of offering public...
charging opportunities and to determine where to site charging stations subject to vehicle travel range constraints.

References
http://dx.doi.org/10.1016/j.trc.2014.02.007
How Should Urban Planners Be Trained to Handle Big Data?

Steven P. French, PhD., FAICP
Dean and John Portman Chair
College of Architecture
Georgia Institute of Technology
Atlanta, GA 30332-0155
steve.french@coa.gatech.edu

Camille Barchers
Graduate Student
School of City and Regional Planning
Georgia Institute of Technology
Atlanta, GA 30332-0155
cbarchers3@gatech.edu

Wenwen Zhang
Graduate Student
School of City and Regional Planning
Georgia Institute of Technology
Atlanta, GA 30332-0155
wunwunchang@gmail.com

July 15, 2014
How Should Urban Planners Be Trained to Handle Big Data?

Steven P. French, Camille Barchers, and Wenwen Zhang

School of City and Regional Planning
Georgia Institute of Technology
Atlanta, GA  30332-0155

Keywords:   Urban Planning, Analytics, Education, Visualization, Simulation

Abstract: Historically urban planners have been educated and trained to work in a data poor environment. Urban planning students take courses in statistics, survey research and projection and estimation that are designed to fill in the gaps in a data poor environment. For decades they have learned how to use the census data, which is comprehensive on several basic variables, but is only done once per decade so is almost always out of date. More detailed characteristics are based on a sample and are only available in aggregated form for larger geographic areas.

But new data sources, including distributed sensors, infrastructure monitoring, remote sensing, social media and cell phone tracking records, can provide much more detailed, individual, real time data at disaggregated levels that can be used at a variety of scales. We have entered a data rich environment, where we can have data on systems and behaviors across smaller time increments and with a greater number of observations on a greater number of factors (Few, 2009; Lohr, 2012). Planners are still
being trained in methods that are suitable for a data poor environment (Contant & Forkenbrock, 1986; Cuzzocrea, Song, & Davis, 2011; Kaufman & Simons, 1995). In this paper we suggest that visualization, modeling, data mining and machine learning are the appropriate tools to use in this new environment and we discuss how planning education can adapt to this new data rich landscape. We will discuss how these methods can be integrated into the planning curriculum as well as planning practice.
Planning methods have been the source of much discussion over the past few decades. Practitioners and researchers have examined what methods planning schools teach and how these are used in practice. The suite of traditional methods courses taught in planning programs – inferential statistics, economic cost-benefit analysis, sampling, and research design for policy evaluation-- remains largely stagnant, despite a rapidly changing reality in which planners are expected to work. Although the focus of this paper is on the impact of big data for planning methods, other variables have also contributed to the need for additional methods to tackle planning problems. The rise of ubiquitous computing and a hyper-connected communication network as well as new private investment in data collection have created an environment in which greater amounts of data exist than ever before. The ability of the planner to analyze and use this data is no longer limited by computing power or the cost of data collection, but by the knowledge that planners possess on data analytics and visualization techniques.

Educating planners with skills that are useful for practice has been a key tenant of many planning programs over the years. Several studies have been conducted to understand better how planning programs succeed or not at this goal. Surprisingly, the most recent comprehensive investigation of planning education and skills demanded by practitioners was conducted in 1986. In this survey, four important conclusions are identified as relevant to how planners were being educated and the professional skills they would be required to use (Contant & Forkenbrock, 1986). They found that the methods taught in planning programs remained highly relevant to the methods needed for practicing planners, and the authors concluded based on their survey results that planning educators were adequately preparing their students to solve planning problems in practice. They cited communication skills (writing and speaking) and analysis and research
design as critical components of planning education and practice, but noted that educators needed to remain vigilant on seeking relevance (Contant & Forkenbrock, 1986). The piece also identified several changes occurring throughout the 1980s that affected the planning profession—the rise of micro-computing and the expansion of methods being offered by planning schools. Contant and Forkenbrock (1986) wrote “…there is little to suggest that planning schools are overemphasizing analytic methods, nor do they appear to be failing to any real extent in meeting the demands of practitioners interviewed. While more techniques are required than these practitioners feel that all planners should understand, it certainly is arguable that this situation is not at all bad.” Thatsurvey of methods is now almost thirty years old, and new realities exist that require educators to expand the methods taught in planning schools (Goodspeed, 2012; Sawicki & Craig, 1996). Yet, despite wide acknowledgement of the changing data landscape, planning curricula still resemble their traditional form. Kaufman and Simons completed a follow-up to this investigation which surveyed planning programs specifically on methods and research design. The more limited focus on this 1995 study “revealed a rather surprising lack of responsiveness among planning programs over time to practitioner demand for [quantitative research methods]” and that “planning programs do not seem to teach what practitioners practice, and not even what practitioners should practice” (Kaufman & Simons, 1995). In a 2002 study focused on the use of technology within planning programs, Urey claims that the haphazard approach with which planning programs have introduced the use of technology to serve larger goals (research, analysis, modeling) might be problematic as increased microcomputing power becomes more widespread. While manual techniques serve learning objectives within planning methods courses, the use of technology is now required (Urey, 2002). This leaves planning educators today with two questions relevant to big data and methods: what
new methods must we now include in our curriculum, and what technology must students become familiar with to employ these methods in an ethical, accurate, and precise way? Given these questions, we reviewed current methods requirements at planning schools to assess whether or not planning programs have begun to respond to these questions and adapt to the changing data landscape.

In a non-scientific review of methods taught at the top ten planning schools (as listed by Planetizen [http://www.planetizen.com/education/planning]), we discovered that almost all programs require that planners be trained in statistics, economic cost-benefit analysis, and research design. Of the programs reviewed, including MIT, Cornell, Rutgers, UC Berkley, University of Illinois Urbana Champaign, UNC Chapel Hill, University of Southern California, Georgia Institute of Technology, UCLA, and University of Pennsylvania, none required students to seek additional data analysis courses outside of the department. Although the review of these programs was not scientific and limited to information published online for prospective students, it does generally persuade us that planning education has yet to see value in teaching planners methods largely reserved for those in the fields of computer science and engineering. We argue, as Contant and Forkenbrok argued thirty years ago, that relevance of planning programs to planning practice is important. Contant and Forkenbrok reminded educators to be vigilant in their understanding of skills that are in demand for practitioners—yet we have failed to do this in regards to our methods curricula.

Big data, although currently a popular topic, is not new—and the concept of big data dates back to 2001, when industry analyst Doug Laney articulated the definition of big data as any data set
that was characterized by the three Vs: Volume, Velocity and Variety (Laney, 2001). Big data sets are characterized by containing a large number of observations, streaming and fast speed and requiring real time analytics. Big data sets are also usually mixed format, joined by a common denominator, but not all of the same type. In sum, any data sets that are too large and complex to process using conventional data processing applications can be defined as big data.

Several pioneers in the industry have already started to process and analyze big data (Lohr, 2012). For instance, UPS now tracks 16.3 million packages per day for 8.8 million customers, with an average of 39.5 million tracking requests from customers per day. The company stores more than 16 petabytes of data. Through analyzing those datasets, UPS is able to identify real time on-road traffic conditions, daily package distribution patterns and together with the latest real time GIS mapping technology, the company is able to optimize the daily routes for freight. With all the information from big data, UPS has already achieved savings in 2011 of more than 8.4 million gallons of fuel by cutting 85 million miles off of daily routes (Davenport & Dyché, 2013). IBM teamed up with researchers from the health care field to use big data to predict outbreaks of dengue fever and malaria (Schneider, 2013). It seems that big data, together with advanced analysis and visualization tools, can help people from a wide variety of industries explore large, complex data sets and extract patterns that were once very difficult to unveil. Given the increasing use of big data across fields that share interests with the field of city planning, planners should more deliberately explore and develop methods for using big data to develop insights about cities, transportation patterns and the basics of urban metabolism.

Data analytics, as a powerful tool to investigate big data, is becoming an interdisciplinary field. There are new programs at universities across the United States that aim to teach students how to
grapple with big data and analyze it using various tools. For this paper, we collected and
reviewed some common tools and skills that are taught in data analytics courses. We gathered
course information from John Hopkins, Massachusetts Institute of Technology, University of
Washington, and Georgia Institute of Technology. We noted that machine learning/data mining
and data visualization are the tools that are frequently taught in the programs to prepare students
to handle big data and some of them are actually quite new to urban planners.

Machine learning is a core subarea of artificial intelligence. Machine learning uses computer
algorithms to create explanatory models. There are different types of learning approaches,
including supervised learning, unsupervised learning, and reinforcement learning. Although
some of the terminologies may be completely new to planners, the actual methods turn out to be
quite familiar. For example, the regression model is one of the methods that is frequently used in
supervised learning process. Planners who work with remote sensing images often apply
supervised classification methods to reclassify the images into land cover images based on
various bands of the image. However, planners may not be familiar with other machine learning
methodologies or algorithms, such as unsupervised learning and reinforcement learning.
Unsupervised learning tries to identify regularities (or clusters or groupings) in the input datasets
without correct output values provided by the supervisors. Reinforcement learning is primarily
used in applications where the output of the system is a sequences of actions (e.g. playing chess).
In this case, what’s important is not a single action, but a sequence of actions that will achieve
the ultimate goal. If the machine learning methods are applied to large databases such as big
data, then it is called data mining. Data mining tries to identify and construct a simple model
with high predictive accuracy, based on the large volume of data. The model is then applied to predict future values.

Most of the programs we reviewed also include data visualization components to help data analysts better communicate their final data products. Some data visualization techniques, such as multivariate data representations, table and graph designs are quite conventional. However, those techniques may also be applied in innovative ways to help convey information behind data in a clearer manner. One example is the information graphics or infographics, which improve human cognition by utilizing graphics to improve the visual system’s ability to extract patterns and trends (Few, 2009; Smiciklas, 2012). The latest trend in data visualization is to take the advantage of webs to present data in an interactive way. To effectively present big data interactively, the designer needs to be equipped with knowledge regarding how human beings interact with computers, and how different interaction types (i.e. filtering, zooming, linking, and brushing) will affect human being’s cognition ability.

In addition to the core courses, these new interdisciplinary programs require the students to master at least one program or query language. SQL is a popular requisite and, in a survey on tools for data scientists, over 71% of respondents used SQL (King, and Magoulas, 2013). Some programs also require students to understand and use open software such as R and R studio.

While these methods for analyzing data may seem completely out of place within a planning methods framework, they actively seek to create ways in which researchers can describe, explore, and explain data. These categories of data analysis are described in depth in Earl Babbie’s *Survey Research Methods*. This text serves as one of many fundamental introductions to methods for planners, and by grouping the new suite of tools available to planners and data scientists within these categories, planners can see how these tools might be useful to them. For
example, data visualization is one of the key ways in which data scientists are exploring big data
sets (Few, 2009). Data visualization acknowledges that our typical methods of data exploration
(descriptive statistics, graphing, and the like) are ill equipped to handle larger data sets, and even
less equipped to communicate information from those data sets to the public and to decision
makers. By introducing planners to the growing field of data visualization, we can expand their
ability to not only to use larger data sets but to communicate the information garnered from those
data sets. As the basis for research, exploration of data sets will allow planners to ask additional
questions. These additional questions will require explanatory analysis, and within this group of
methods, tools such as machine learning and data mining will help planners generate predictive
models from larger data sets.

Many data sets that planners have to deal with in the future may turn out to be big data. Credit
card data or web browsing histories may help planners to predict the focus of emerging public
concerns. As a matter of fact in MIT’s big data courses, there is a case study regarding how to
utilize the Google search records to estimate the trends within the real estate industry (MIT,
2014). Social media, such as Twitter and Facebook, have already become powerful information
sources regarding almost every aspect of social life. Analysis of twitter feeds can help to identify
the extent and intensity of hazard events. There are already studies towards how to utilize
information extracted from Facebook’s friend list to forecast the use of airplanes. GPS or real
time transportation information can help planners to calibrate and develop more accurate activity
based travel demand models to forecast future travel patterns. Moreover, the real time
information about energy flows such as water, sewer, and electricity flows may equip planners
with critical information to design more energy efficient and sustainable cities to make built
environment more resilient to natural hazards and climate change. Planning is defined by our special affinity for place based issues, and this focus on place will be one of the critical ways in which typical data sets can become “big data.” Location is the ultimate relational field, and our ability to link data sets through location will create big data sets that are especially useful to planners. If location is the ultimate relational tool, than planning data sets will only continue to increase in size, speed, and format in the future. The importance of teaching planners how to accurately and precisely examine and explore this data cannot be understated, yet, our work to prepare this paper leads us to believe that planning programs have not yet taken the steps required to introduce these methods to planning students.

The big data analysis tools, such as machine learning and data visualization, can help planners to make better use of the big data sets. The Memphis Police Department used machine learning and data mining approaches to predict potential crime based on past crime big data, the serious crime rate was reduced by approximately 30%. The city of Portland, Oregon optimized their traffic signals based on big traffic data, and was able to reduce more than 157,000 metric tons of CO$_2$ emissions in six years (Hinssen, 2012). In sum, the machine learning techniques can help planners to analyze the future development of urban areas in a more accurate way to eliminate or at least ease some the current urban problems. The explanatory power of machine learning will be critical for planners seeking to use big data to solve long-term challenges in cities and communities.

Data visualization has always been considered useful in the planning process, primarily as a communication method. However, it is now critical in exploring large data sets. Data visualization can help planners better understand how people live, work and behave within urban
context. When paired with more explanatory tools such as machine learning, data visualization becomes a critical step in the planning process. Visualization can also continue to be used as a way for planners to convey their planning concepts to corresponding stakeholders during the public participation process. In this way, visualization is used as an interpretation toolkit to help people digest the complex analysis results from big data. Planners continue to be more comfortable using traditional graphs, tables, and animation images to visualize their results. However, some planners are now using more advanced web based tools to display the information in interactive ways to encourage public participation. This trend has been on the rise for some time, and the demand for practitioners with visualization skills continues to increase (Few, 2009; Goodspeed, 2012; Sawicki & Craig, 1996).

We argue in this paper that planners would benefit greatly from the introduction of more advanced methods of descriptive, exploratory, and explanatory data analysis in order to more effectively use an ever increasing amount of relevant data. However, some key questions for additional research emerged as we prepared this paper. Two characterizations of planning prove to be challenging when introducing new methods to the planning student and practicing planner—the field of planning is inherently place based, and as such it has the potential to take typical planning data and transform it into big data by introducing volume and mixed format information into databases. Secondly, the field of planning is concerned with the long-term. Thus far, big data has been analyzed to provide insights into short term challenges—the speed that is required to transform a data set into big data means that we are frequently examining real time data at such minute discrete time fragments that it verges on becoming continuous. As planners, we need to be asking a larger question that relates to not just what methods can be used to employ this data, but how this data can be employed in our search for long-term solutions.
does the use of minute by minute Twitter text analysis related to planning issues allow us to reframe planning issues for years to come? How does second to second transportation data give us a new way to understand transportation patterns that matter beyond the day and into the next decade? Although we did not set out to answer these questions in this paper, we do believe that posing them along with giving planners insight into how we can use data that is becoming increasingly large, fast moving, and varied, will help us redefine planning methods for the next generation of planning students and practitioners. Big data is an exciting new asset for planners who have always struggled to explore and explain patterns and trends based on limited observations of discrete data—we should make the best use of this data by giving our planners the tools with which to analyze it and communicate it. Like others who have written on the topic of big data in cities, we do caution that data should not be used or analyzed purely for data’s sake, that is, planners are tasked with a more nuanced goal than computer scientists: they must find ways in which to use the data to make existing communities better and to provide better solutions than were previously available (Mattern, 2013; Sawicki & Craig, 1996). In order to help planners achieve these goals, we must revamp our methods offerings in our planning programs to better represent the ways in which we are interacting with big data as researchers.
References


Big Spatio-temporal Network Data Analytics for Smart Cities: Research Needs

Venkata M.V. Gunturi, Shashi Shekhar, William Northrop, Andrew Kotz

a Department of Computer Science and Engineering, University of Minnesota, Minneapolis, USA
b Department of Mechanical Engineering, University of Minnesota, Minneapolis, USA

{gunturi, shekhar}@cs.umn.edu, {wnorthro, kotzx029}@umn.edu

Abstract: Given, a collection of emerging valuable spatio-temporal network (STN) datasets, e.g., temporally detailed roadmaps, GPS tracks, traffic signal timings etc., and smart-city use-cases; the problem of Big-STN analytics is concerned with extracting valuable information from Big-STN data using modern computing platforms. Big-STN analytics can add significant value in several smart-city applications such as, urban navigational systems, urban transportation management and planning. However, the envisaged Big-STN analytics pose significant challenges for current state-of-the-art techniques. First, a semantic gap is raised due to mismatch between nature property recorded in Big-STN data and capabilities of current data-models. For instance, current data-models based on binary relations cannot represent holistic properties, e.g., travel-time or idling experienced by an urban traveler on a route with series of coordinated traffic signals. Second, the linear nature of the Big-STN patterns raise significant semantic challenges for the current spatial pattern techniques which are mostly focused on geometric shapes, e.g., circles and rectangles. The objective of this article is to explore Big-STN analytics along two dimensions: (a) explore the use-cases related to smart cities which can be enhanced and, (b) elucidate the research needs in terms of new data-models and pattern families for these use-cases.

Keywords: Spatial Databases, Spatial Data Mining, Road networks, Routing, Spatio-temporal networks.

1. Introduction

Smart cities can be considered as an "interconnected eco-system" thriving through a diverse set networks for a collective sustenance, communication, distribution and delivery for its entities. Examples of such networks include, transportation network (for distribution and delivery); water, electricity, gas and sewage-disposal networks (for sustenance); internet and tele-communication networks (for communication) (Kelmelis & Loomer, 2003). A common characteristic among all these networks is that they are either physically embedded in space (e.g. transportation, water, sewage-disposal networks) or, use space for increasing their reach (e.g. transmission towers and repeaters for electricity and tele-communication networks).
Increasingly, a large number of urban sensors in major cities (Boyle, Yates, & Yeatman, 2013) have started to produce a variety of datasets representing both historic and evolving characteristics of these smart-city networks. Each of these datasets record a certain property or a phenomena on the smart-city network spread over space and time. A collection of such datasets is referred to as Big Spatio-Temporal Network (Big-STN) data. On transportation network alone we have emerging datasets such as, temporally detailed (TD) roadmaps (S Shekhar, Gunturi, Evans, & Yang, 2012) that provide travel-time for every road-segment at 5 minute granularity, traffic signal timings and coordination (H. Liu & Hu, 2013.), GPS tracks annotated with engine measurements data (e.g., fuel consumption) (S Shekhar et al., 2012; Jing Yuan et al., 2010), traffic information from vehicle-to-vehicle communications (Lee, Magistretti, Gerla, Bellavista, & Corradi, 2009; Wolfson & Xu, 2014). Other sample Big-STN datasets include, meter readings from water (Yang, Shekhar, Dai, Sahu, & Naphade, 2011) and energy distribution (electricity and gas) (Boyle et al., 2013) networks as well as traffic information from communication networks (internet and telecom). Due to space limitation, this paper focuses only on Big-STN data generated from transportation network.

Given such an instance of Big-STN data and a set of smart city use-cases, the objective is to probe into a realization of what we refer to as Big-STN analytics. This realization includes two aspects: (a) Big-STN query processing and, (b) Big-STN mining.

Value of Big-STN Analytics: Realization of Big-STN analytics is important due to its value addition potential in several smart-city application use-cases. Examples include urban navigation, transportation infrastructure management and planning. Table 1 provides few potential extensions (enabled by Big-STN analytics) of the same traditional routing query "what is the shortest route between UMN and Airport?" on a typical urban transportation network. As the table shows, these extensions can be realized along the dimensions, (a) Big-STN query processing and, (b) Big-STN mining. The extensions enabled by Big-STN query processing come through emerging features being increasingly captured in Big-STN data. For example, historical congestion information available at multiple departures times (from TD roadmaps) allows us to compare different candidate routes for a range of potential departure-times (Q1 in Table 1). Similarly, traffic signal timing information presents an opportunity to make a route choice not only based on typical travel-time information, but also to incorporate typical idling
experienced at red lights (Q2). And lastly, through GPS traces, we could suggest more personalized routes (Q3 in Table 1) or something based on what typical commuters prefer (Q4 and Q5 in Table 1).

Table 1: Sample extensions of the traditional routing query as realized by Big-STN Analytics.

<table>
<thead>
<tr>
<th>Big-STN Query Processing</th>
<th>Big-STN Mining</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1. What is the shortest route between UMN and Airport over 7:00--12:00noon?</td>
<td>P1. What is a representative set of routes usually taken by commuters to Airport?</td>
</tr>
<tr>
<td>Q2. What is the typical delay experienced on Hiawatha route to Airport?</td>
<td>P2. What portion of these routes show low fuel economy or higher greenhouse gas emissions?</td>
</tr>
<tr>
<td>Q3. What routes are most fuel efficient for my driving behavior?</td>
<td>P3. What is a representative driving behavior that achieves best fuel economy on the Hiawatha route to Airport?</td>
</tr>
<tr>
<td>Q4. What is the common route taken by commuters to Airport from UMN?</td>
<td>P4. In what kind of a journey (acceleration and braking patterns) does a vehicle experience poor fuel economy over others?</td>
</tr>
<tr>
<td>Q5. What is the emerging preferred route to Airport due to construction?</td>
<td>P5. What is a hotspot of such patterns?</td>
</tr>
</tbody>
</table>

The Big-STN mining extension of the above mentioned traditional routing query could focus on extracting non-trivial information from the Big-STN data. P1, P2, P3 and P4 listed in Table 1 are some candidate questions which can be answered through Big-STN mining. Answers to these questions could not only help in providing a richer route recommendation for travelers and logistic companies, but also help traffic management agencies to plan the transportation infrastructure better. For example, representative routes returned by P1 could give us an insight into common choices of the commuters. Combined with P2, one could probe into the possibility of upgrading the road infrastructure (e.g. add a new lane, coordinating traffic signals etc.) towards a broader goal of city wide fuel efficiency and reduced greenhouse gas emissions. Similarly logistics companies could use the information returned from P3, P4 and P5 to fine tune their engines (e.g. adjust spark timing or rate of fuel injection) to achieve better fuel efficiency for their fleets.
Potential of Big-STN analytics was also highlighted in recent McKinsey report which estimated savings of "about $600 billion annually by 2020" in terms of fuel and time saved (Manyika & others, 2011.) by helping vehicles avoid congestion and reduce idling at red lights or left turns. Similarly, preliminary evidence of its value proposition can also be seen through experience of UPS, which saved millions of gallons of fuel annually by exploiting such ideas (Davenport & Dyche, 2013; Esposti, 2014; Lovell, 2007).

**Organization of rest of the paper:** The rest of the paper is organized as follows: Section 2 describes the Big-STN datasets considered in this paper. An interpretation model of these datasets is given in Section 3. We detail our proposed vision of Big-STN analytics in Section 4 and highlight the challenges it poses to the current state-of-the-art. Finally, we conclude in Section 5 with a brief look at the future work.

![Figure 1: Sample Transportation network for illustrating Big-STN data (best in color)](image)

### 2. What is Big-Spatio-Temporal Network Data?

A *Spatio-temporal network (STN)* dataset can be defined as a collection of entities in a networked system which has certain spatial and temporal characteristics. The spatial characteristics arise due to a natural embedding of the entities inside a space. Whereas, the temporal features of the dataset arise due to temporal variation of certain attributes of the entities. A collection of such datasets for the same networked system is referred to as *Big-STN* data. Such networks arise in several domains including, transportation, fluid flow in river networks, traffic flow in computer networks etc. In this paper, we limit to transportation network as our desired...
networked system. Figure 1 shows a sample transportation network in Minneapolis area (to the left). On the right, it shows its simplified version where arrows represent road segments and labels (in circles) represent an intersection between two road segments. Locations of traffic signals are also annotated in the figure.

Illustrative Big Spatio-temporal Network Data: We consider an instance of Big-STN data on our sample networked system shown in Figure 1 by considering three types of STN datasets: (a) temporally detailed (TD) roadmaps (S Shekhar et al., 2012), (b) traffic signal data (H. Liu & Hu, 2013.) and (c) annotated GPS traces (S Shekhar et al., 2012; Jing Yuan et al., 2010). Each of these measures either historical or evolving aspects of certain travel related phenomena on our system. TD roadmaps store historical travel-time on road segments for several departure-times (at 5min interval) in typical week. For simplicity, essence of TD roadmaps is illustrated in Figure 2 by highlighting the morning (7:00am--11:00am) travel time only on segments A-F, F-D and S-A (7mins, 11mins and 9mins are respectively). The travel-times of other road segments in the figure (shown next to arrows representing roads) are assumed to remain constant. The figure also shows the traffic signal delays during the 7:00am--11:00am period. Additionally, the traffic signals SG1, SG2, SG3 are coordinated such that in a journey towards D (from S within certain speed-limits), one would typically wait only at SG1. Similarly, a journey towards D starting on segment B-C (after SG1) would have to wait only at SG2.

Map-matched and pre-processed (Y. Zheng & Zhou, 2011) GPS traces are another component STN dataset in our Big-STN data. Each trace in this dataset consists of a sequence of road-segments traversed in the journey along with its schedule denoting the exact time when the
traversal of a particular segment began (via map-matching the GPS points). GPS traces can potentially capture the evolving aspect of our system. For instance, if segment E-D in Figure 2 is congested due to an event (a non-equilibrium phenomena), travel-time denoted by TD roadmaps may not be accurate anymore. In such a case, one may prefer to follow another route (say C-F-D) which other commuters may be taking to reach D.

Additionally, GPS traces can also be annotated with data from engine electronic control computers to get richer datasets illustrating fuel economy of the route. We refer to them as annotated GPS traces. Figure 3 illustrates a sample annotated trace highlighting the variation in multiple engine sub-system parameters like engine speed in revolutions per minute, fueling rate and vehicle speed along a route in St. Paul, MN. These parameters can give us an insight into the driving style of the person driving the vehicle for a given roadway and traffic condition. For example, a highly variable engine speed could denote more aggressive, short but frequent periods of acceleration and braking by the driver.

3. Interpretation Model of Big-STN Data

Information recorded in Big-STN data usually corresponds to a particular exemplary behavior considered by domain experts while putting together the dataset. Interpreting the datasets in terms of their intended exemplary behavior is crucial. We first describe the common exemplary behaviors considered followed by their effect on travel-time and fuel consumption, common properties recorded in Big-STN data.

Depending on the dataset, the exemplary behavior could be the experience of a representative individual traveler, a representative platoon of vehicles traveling towards a...
common destination, or the characteristics of an individual commuter. TD roadmaps is a case where the underlying reference behavior of interest was modeling the experience of a representative individual traveler (within speed-limits) on road segments at multiple departure-times in a day of a typical week.

On the other hand, datasets related to traffic signals timings and coordination is put together with a representative platoon-of-vehicles as the underlying exemplary behavior. Here, the cycle-lengths and phase-gap between different signals is set to provide a smooth flow of traffic (a platoon of vehicles) along arterial roads and highways with minimum stops (Koonce & et al., 2008). Such coordination is a common practice in major cities to reduce idling at red lights. For instance, recently about 4400 traffic signals in Los Angeles were coordinated (Lovett, 2013) with a goal of increasing the traffic speeds by about 16%. Here, the cycle-lengths (i.e., duration of one cycle) and phase-gap (time offset between cycles of different signals) are set in such a way that, first a platoon gets collected at intersection and then it gets 'transferred' smoothly through other intersections. For instance, consider a sample journey from S to D via S-B-C-E-D in Figure 2. This journey would traverse road segments S-B, B-C, C-E and E-D. And while doing so it would experience the traffic signals SG1, SG2 and SG3. Figure 4 illustrates the cycle-lengths and phase-gap for these signals. The figure shows that the phase-gap between SG1 and SG2 is set to the typical travel-time on segment B-C (ignoring queue delays). Similarly, the phase-gap between SG2 and SG3 is set to typical travel-time of C-E. Thus, vehicles traveling towards D would typically wait only at SG1 before being transferred through SG2 and SG3 wait-free. In other words, in this journey initial waiting at SG1 renders SG2 and SG3 wait free. If the effects of immediate spatial neighborhood are referred to as local-interactions, e.g. waiting at SG1 delays entry into segment B-C, then this would be referred to as a non-local interaction as SG1 is not in immediate spatial neighborhood of C-E (controlled by SG2) and E-D (controlled by SG3).
And lastly, datasets with annotated GPS traces reflect the choices and experiences of an individual *commuter* (not necessarily a representative individual) of a controlled study or volunteered itineraries (Jing Yuan et al., 2010).

### 3.1 Effect of reference behavior:

Travel-time recorded under different exemplary behaviors will have different characteristics. In the case of datasets where the focus is a representative individual traveler, typical travel-time for a larger instance (e.g., route) when determined by joining the value for smaller instances (e.g., individual segments) or vice-versa would retain its intended reference behavior. We refer to this as *decomposable travel-time* or travel-time behaving like a *decomposable property*.

On other hand, the travel-time in datasets with platoon-of-vehicles or individual-commuter as the underlying exemplary behavior may not always exhibit such decomposability due to non-local interactions. Consider again the sample journey of S-B-C-E-D through a series of coordinated traffic signals in Figure 2. Here, the total travel-time measured on a typical journey from S to D would not include any waiting at intersections C (signal SG2) and E (signal SG2). This is, however, not *true* as a traveler starting at intersection C (or E) would typically wait for some time SG2 (or SG3). We refer to this kind of behavior, where properties (e.g. travel-time) measured over larger instances (e.g. a route) lose their semantic meaning on decomposition, as *holism*. For our previous example, we say that the total travel-time measured over the route S-B-C-E-D was behaving like a *holistic property*.

Apart from the previous example, travel-time (and fuel consumption) experienced by a commuter on a road-segment, as derived from his/her annotated GPS trace, is also holistic in nature. Here, the travel-time (fuel consumption) experienced would depend on his/her initial velocity gained *before* (a non-local interaction) entering the particular road-segment under consideration.
4. Big Spatio-temporal Network Analytics

As mentioned earlier, Big-STN analytics consists of following components: (a) Big-STN query processing and (b) Big-STN Data Mining. We now discuss both of these in detail, providing specific problem definitions, challenges to the state-of-the-art and open research questions.

4.1 Big-STN query processing: This component is concerned with the development of a querying framework over Big-STN data for common use-cases queries. Such a framework would consist of a representational data-model and a suite of query algorithms which can harness it to provide results for a given set of use-case queries. We now begin with a discussion focusing on the representational aspects of Big-STN data followed a list of open research questions.

The problem of modeling Big-STN data can be defined as follows: Given a collection of STN datasets (e.g. TD roadmaps, traffic signal data and annotated GPS traces) and a set of use-case queries, the aim is to build a unified logical data-model across these datasets which can express a variety of travel related concepts explicitly while supporting efficient algorithms for given use-cases. The objective here would be to balance the trade-off between expressiveness and computational efficiency. Such a unified logical data-model would enable richer results by allowing query algorithms to access and compare information from multiple datasets simultaneously.

Modeling a collection of STN datasets is challenging due to the conflicting requirements of expressive power and computational efficiency. In addition, the growing diversity of Big-STN data requires modeling a variety of concepts accurately. For instance, it should be convenient to express our previously described holistic properties in Section 3. Consider again our previous example the journey S-B-C-E-D through a series of coordinated traffic signals SG1, SG2 and SG3 in Figure 2. Now, a candidate model could choose to decompose this journey (from S to D) down to the level of individual road segments, allowing easier storage and efficiency through binary relations. While such a model achieves efficiency it loses the semantic meaning of the data (non-local interactions) in this case and for other holistic properties in general.
Current approaches for modeling STN datasets such as time aggregated graphs (George & Shekhar, 2006, 2007), time expanded graphs (Kaufman & Smith, 1993; Köhler, Langkau, & Skutella, 2002) and, (Güting, 1994; Hoel, Heng, & Honeycutt, 2005) are most convenient when the property being represented can be completely decomposed into properties of binary relations. In other words, they are not suitable for representing the previously described holistic properties. Current related work would represent our previous signal coordination scenario using following two data structures (See Figure 5): one containing travel-time on individual segments (binary relations) S-B, B-C, C-E, and E-D; the second containing the delays and the traffic controlled by signals SG1, SG2, and SG3 (also binary). However, this is not convenient as non-local interactions affecting travel-times on some journeys (e.g. S-B-C-E-D) are not expressed explicitly. Note that this representation would have been good enough if SG1, SG2 and SG3 were not coordinated.

Ideally, the representation model should express the non-local interactions more explicitly. Figure 6 illustrates the spirit of our proposal for the previous signal coordination scenario. Here, we propose to represent the journey as a series of overlapping sub-journeys, each accounting for a non-local interaction. The first entry in the figure corresponds to travel-time experienced on the sub-journey containing road segment S-B (3mins) and delay at SG1 (max delay 90secs). This would be between 3mins and 4mins 30secs (no non-local interactions in this case). Next we would store travel-time experienced on the sub-journey containing road segment S-B (3mins), delay at SG1 (max delay 90secs), segment B-C (8mins) and delay at SG2 (max 90secs) as between 11mins and 12mins 30secs. Notice that we did not consider the delay caused by SG2 due to non-local interaction from SG1. This process continues until all the possible non-local interactions are covered.
**Open Research Questions:** A key task while developing a representational model for Big-STN data would be to balance the tradeoff between computational scalability and accuracy of representation of important travel related concepts such as non-local interactions. Further, the proposed model should provide a seamless integration of TD roadmaps, traffic signal data and, experiences of commuters through GPS traces. This would allow the routing framework to explore candidate travel itineraries across multiple sources to get richer results. For example, routes from TD roadmaps, which recommend based on historic congestion patterns, can be compared (during candidate exploration phase itself rather than offline) against commuter preferences for factors like fuel efficiency and “convenience”. Additionally, one also has to develop a set data-types and operations for queries on this representational model. This would involve addressing questions such as: What would be minimal set of operators expressive enough to represent spatio-temporal routing queries about a traveler’s experience? Do they facilitate the design of dynamic programming based methods? How should the set be refined to address the conflicting needs of expressive power and support for efficient algorithms? And lastly, we would also have to explore the computational structure of the various use-case queries. A sample research question could be: Does the classical dynamic programming for fastest query hold in the face of new features like fuel efficiency, signal delays and travel-time information at multiple departure-times?

**4.2 Big-STN Data Mining:** This section describes our proposed concept of Big-STN data mining through a few representative pattern families and proposes few open research questions for further investigation. We define Big-STN data mining as a set of data mining techniques which focus on physical concepts while defining the patterns of interest. These patterns in their simplest form can be defined as spatio-temporal patterns whose semantics are enhanced through physical science-based concepts. These concepts implicitly dictate a “sub-space” defined in terms of a certain relationship among the physical dimensions where the patterns can be expected to reside. As a simple example, consider the dimensions of space, time, and vehicle-speed available as features in our annotated GPS traces dataset. In general this relationship is likely to be defined over many engine related dimensions including features like, engine-speed, acceleration, fueling-rate, etc.
A physics-agnostic pattern defined on these dimensions would consider each one of them separately to account their effect on the fuel economy. Some patterns include: What are regions (space only) are likely to show poor fuel economy? What times of day do we typically observe low fuel economy (time only)? How is engine speed correlated with fuel consumption (only engine data)?

In contrast, a Big-STN pattern incorporates a relationship among these dimensions to obtain richer semantics. A sample relationship is “as experienced by the engine” while driving on a real road network. Notice that such a relationship is defined in terms of all of the previous dimensions and perhaps including some additional ones such as velocity and fueling rate. Such a relationship gives us a richer set of patterns. Examples include: In what kind of a journey (space, time and vehicle-speed dimensions) does a vehicle experience poor fuel economy over others? What is a “hotspot” (in terms of space, time and vehicle-speed dimensions) of such patterns? What is a representative drive cycle (a journey in space, time, vehicle-speed and other engine related dimensions) which achieves best (or worst) fuel economy in a downtown region? What engine-related parameters correlate with each other under what space-time dimensions?

![Route A](a) Conventional vehicle ![Route B](b) Hybrid vehicle ![Route A](c) Conventional vehicle ![Route B](d) Hybrid vehicle

**Figure 7:** Fueling rate (red=high, green=low) on two vehicles observed on two routes at same time-of-the-day

Figure 7 illustrates fueling rate hotspot patterns on a transportation network designating area of high fuel use for conventional and hybrid vehicles using measured engine data from buses in Minneapolis-St Paul area. While some hotspots may be common across conventional and hybrid vehicles, Figure 7 shows that other hotspots are unique to vehicle types. Finding such specialized hotspots of inefficiency could spur next level of research questions such as: What is the scientific explanation behind a certain low fuel economy observation? Was it due to a particular acceleration-breaking pattern? Can this pattern be coded into the engine control unit for better fuel economy?
Mining Big-STN patterns poses both semantic and computational challenges. Semantically, traditional techniques for mining statistically significant spatial patterns (Kulldorff, 1997a) focus on geometric shapes, e.g., circles and rectangles. As a consequence, when forced on to the space of transportation network, they would have lower values of likelihood ratios and statistical significance (e.g., p-values) and thus may lose several linear shaped patterns which might be interesting. This is illustrated in Figure 8 using transportation safety dataset about pedestrian fatalities explored in our preliminary work (Oliver et al., 2014) with transportation network aware patterns. Pedestrian fatalities usually occur on streets, particularly along arterial roadways (Ernst, Lang, & Davis, 2011). However, the results generated by geometric shape based approach (Kulldorff, 1997a) does not capture this. From Figure 8(a), it is clear that the circle-based output is meant for areas, not streets. In contrast, a non-geometric-shape based approach focusing on semantics of transportation network (Figure 8(b)) through constructs like routes can identify four colored road-stretches with better p-values.

![Figure 8](image_url)

(a) Geometric-shape based technique (circles)  
(b) transportation network aware technique (routes)

Figure 8: Statistically significant patterns discovered by geometric-shape based approach and transportation network aware approach.

However, transportation network based non-geometric approaches face significant computational challenges in the face of Big-STN patterns due to following two reasons. First, set of candidate road-stretches is very large (exponential in number of road-intersections in roadmap) even for a single engine-dimension (e.g., fuel efficiency) or real-world-dimension property (e.g., traffic). Second, the candidate space grows even bigger for multiple dimensions related to engines, vehicles and real-world context. We now detail few research questions for addressing the previously mentioned challenges for a hotspot or summarization pattern.
**Open Research Questions:** As a first step towards realizing Big-STN mining, we need to conceptualize the Big-STN patterns and develop suitable interest measures. The objective here would be to explore the tradeoff between the semantic meaning and computational scalability of the precise mathematical quantification of the pattern. For example, a common task in both hotspot detection and summarization is to define a distance metric between two journeys (in terms of all its dimensions in the dataset). In order to obtain a semantically meaningful and computationally efficient distance metric one may plan to explore research questions such as: What properties should the distance measure possess? Should it demonstrate triangle inequality (e.g. shortest distance)? Should it be amenable to dynamic programming based approaches? How suitable it is to construct a meta-level interest measure (e.g. a likelihood ratio) from this metric for statistical significance?

After conceptualizing the patterns, we need to investigate scalable computational techniques for computing them. This is challenging due simultaneous requirements for ability to scale-up to large datasets and also maintain a formal characterization (in terms of correctness and completeness) of the solution quality. Additionally, it should also ensure statistical significance of the discovered pattern through techniques like p-value and Monte-Carlo simulations. A typical sub-task for both hotspot and summarization patterns involves computing pair-wise distances computation among all pairs of journeys. Some sample research questions for this could be: Can we compute the distance metric using standard graph traversal algorithms? Can they be coupled with a dynamic programming based approach (depending on the nature of distance metric)? Additionally, one can explore other ideas like creating a “virtual super journey” from which the exploration could start. This would help in computing distances between several pairs in single iteration.

**Other Related work in STN mining:** Current literature towards big data analytical techniques in the form of spatial data analysis techniques such as spatial statistics (Anselin & Getis, 2010; Kulldorff, 1997b; Neill & Moore, 2004) and spatial data mining (Huang, Shekhar, & Xiong, 2004.; S Shekhar, Zhang, Raju, & Huang, 2003; Shashi Shekhar, Evans, Kang, & Mohan, 2011; Yoo & Shekhar, 2006.) have explored patterns in the geographic space, with some patterns families defined over few time snapshots (Mohan, Shekhar, Shine, & Rogers, 2010, 2012). These patterns mostly conform to geometric shapes or structure created out of neighborhood relation in
a geographic space. In other words, they do not model the semantics of a transportation network which constraints a pattern to a linear space (e.g., a path). Other related work in this area includes work on GPS track mining (Bastani, Xie, Huang, & Powell, 2011; Doytsher, Galon, & Kanza, 2011; Fawcett & Robinson, 2000; Fu, Hu, & Tan, 2005; Gonzalez, Han, Li, Myslinska, & Sondag, 2007; Horvitz & Krumm, 2012; Hsieh, Li, & Lin, 2012; Krumm, 2011; Letchner, Krumm, & Horvitz, 2006; X. Li, Han, Lee, & Gonzalez, 2007; Z. Li et al., 2010; B. Liu, 1997; W. Liu, Zheng, Chawla, Yuan, & Xing, 2011; McGinty & Smyth, 2000; Min & Wynter, 2011; Sacharidis et al., 2008; Wei, Zheng, & Peng, 2012; Won, Kim, Baek, & Lee, 2009; J Yuan, Zheng, Xie, & Sun, 2011; Jing Yuan et al., 2010; Zhang et al., 2011; K. Zheng, Zheng, Xie, & Zhou, 2012; Y. Zheng & Zhou, 2011) which mine only the space-time locations of the vehicles to improve route recommendation and estimate traffic congestion patterns in a city. In summary, the current techniques when applied to Big-STN data would often result in patterns which are hard to interpret due to either their in-ability to model the transportation network semantics or limiting to only space-time locations when additional physics parameters (e.g. engine speed in revolutions per minute) need to be considered.

5. Conclusion

Big-STN data analytics is important for societal applications such as urban navigational systems and transportation management and planning. It is however challenging to do so due to following reasons: First, holistic properties increasingly being captured in the Big-STN data raise representational challenges for the current state-of-the-art data-models for spatio-temporal networks. And second, the linear nature of the Big-STN patterns raise significant semantic challenges to the current state-of-the-art in the area of mining statistically significant spatial patterns which mostly focus on geometric shapes, e.g., circles and rectangles. In future, we plan to explore the research questions in the paper towards formalizing the notion of Big-STN data analytics. Specifically, we plan to build representational models and computational techniques which attempt to address the aforementioned challenges of Big-STN analytics.
Acknowledgment: This work was supported by: NSF IIS-1320580 and 0940818; USDOD HM1582-08-1-0017 and HM0210-13-1-0005; IDF from UMN. We would also like to extend our thanks to Kim Koffolt for improving the readability of this paper. The content does not necessarily reflect the position or the policy of the government and no official endorsement should be inferred.

References


I take the view of a city as a platform. As a platform, a city has infrastructure, processes, organizations, individuals, and technology as components. Additionally, cities are comprised of technical (e.g. sensors), social (e.g. humans), and socio-technical components (e.g. processes). The glue that holds these components together and enables integration and coordination to occur is data and information. The effective and efficient management of information is not only critical to ensure that each of the components operate optimally but also ensures that the overall system, the city, achieves its overall objectives. In this paper, I focus on three key data dimensions in the context of urban informatics: big, open, and mobile data. Key issues within each data dimension are presented. The paper builds on several research projects on smart cities, urban informatics, and policy informatics. Data collected during these projects includes over 45 case studies, over 60 interviews with key informants, analysis of over several thousand pages secondary data, and an examination of over 70 technology solutions that span mobile apps, online crowdsourcing platforms, sensors, analytical and visualization technologies, and associated urban technologies. The paper puts forth several considerations that need to be accounted for when discussing the potential of data and technologies to transform our urban spaces towards the goals of making them intelligent, livable, sustainable, and resilient.

**Keywords:** Urban Informatics; Big Data; Mobile Data; Open Data; Civic Technologies; Crowdsourcing; Smart Cities; Resilience; Intelligent Cities
Introduction

Cities are complex systems (Batty, 2007). Complexity arises from the fact that cities are comprised of multiple systems, sub-systems, and components that interact with each other in both predictable and unpredictable manners giving rise to emergent outcomes. Traditional mechanisms to governing, designing, and planning mainly center on top-down approaches and fail in the face of increasing complexity. Innovative, bottom-up approaches are needed to capture real-time data on the environment, fuse data to form information that can be leveraged towards decision-making in real-time, and opportunistically deal with challenges. As we continue to see the role of cities expand and the number of megacities continue to rise, there remains much interest to increase the intelligence of our cities, i.e. make them smarter (Desouza, 2012; Neirotti et al., 2010).

Making a city smarter, or building a new smart city from scratch, is not an easy feat. Implementation of urban informatics is critical for achieving this goal (Foth, Choi, and Satchell, 2011, Desouza 2012, Bays and Callen, 2012). According to Foth et al. (2011), “urban informatics is the study, design, and practice of urban experiences across different urban contexts that are created by new opportunities for real-time, ubiquitous technology and the augmentation that mediates the physical and digital layers of people networks and urban infrastructures” (p. 4). By applying urban informatics, cities can: (1) harness existing data to improve service delivery, (2) construct new databases to increase the effectiveness of operational and planning decisions, and (3) increase citizen engagement (Bays and Callen, 2012). Urban informatics as a field of study has a rich history (see Foth, 2009; Mitchell, 1995; Purcell, 2006; Townsend, 2000).

We are currently witnessing a resurgence when it comes to interest in urban informatics due to two key developments. First, the diffusion of information and communication technologies (ICTs) makes it possible for the masses to interact with data in real-time. The explosion of mobile technology is reshaping communities around the world. In January 2014, 98% of adults in the US owned a cell phone, of which, 58% were smartphones (Pew Research Report, 2014). Moreover, it is estimated that more than 4 billion users worldwide will have access to mobile phones by 2018 (Kearney, 2013). Emergent economies have seen the greatest levels of growth in mobile access in recent years, e.g., the number of persons owing mobile phones in Sub-Saharan Africa increased from 16 million in 2008 to over 500 million in 2012 (Rotberg and Aker, 2013). There is no doubt that number of mobile users have grown across the world, and this number will continue to increase. The access to ICTs is opening up opportunities for us to engage citizens in the planning, designing, and governing of their urban spaces, both in the developed and the developing world (Desouza and Smith, 2014; Desouza 2014c).

Second, interest in increased data accessibility creates opportunities for citizen engagement. Open data movements, which share data about cities and public institutions (agencies), are spurring up all across the globe (Open Data Barometer, 2013; Manyika et al., 2013). Agencies are making data available to the public about all facets of a city from transit to crime. In addition, agencies are liberating data that were traditionally locked up within administrative systems. The overriding goal here is to increase transparency, thereby increasing trust in government while also enabling more collaborative and participatory governance (Obama, 2009). For example, in 2010, when the city of London published public data in a machine-readable format, individuals outside the public sector realm used the data to create services for communal good. One developer created online map of the city of London Tube and another individual created a bike route map for the city. These apps help people better understand the city’s transportation system (Halper, 2011).

In this paper, I take the view of a city as a platform. As a platform, a city has infrastructure, processes, organizations, individuals, and technology as components. Additionally, cities are comprised of technical (e.g. sensors), social (e.g. humans), and socio-technical components (e.g. processes). The glue that holds these components together and enables integration and coordination to occur is data and information. The
effective and efficient management of information is not only critical to ensure that each of the components operate optimally but also ensures that the overall system, the city, achieves its objectives. Today, some of the objectives of a city include being intelligent (smart), livable, sustainable, and resilient. Given that a city is a complex system, it is unreasonable and incorrect to assume that a city can thrive with only top-down flow of data and information. Cities are shaped by the emergent, bottom-up, interactions between components. Top-down planning can be valuable, however, it is inefficient and ineffective when it comes to designing, planning, and governing complex entities. The two trends noted above, the diffusion of ICTs and the provision of data to the public, are enabling cities to engage with citizens in a deeper manner than ever before when it comes to urban planning and governing. The goal of this paper is to highlight critical considerations that still need to be resolved if we are truly take advantage of the increasing diffusion of technology and data deluge to build smarter cities. I focus on these three big trends when it comes to data: big, mobile, and open. I specifically look at how these three types of data play out when it comes to our urban infrastructure and governance towards the design of smarter and more resilient cities. Drawing from over three years of research in smart cities into smart cities and urban informatics (for more information, see – Desouza 2012a, 2012b, 2014a, 2014b, 2014c; Desouza and Bhagwatwar, 2010, 2012a, 2012b, Forthcoming; Desouza and Flanery, 2013; Desouza, 2012, 2014; Desouza and Schilling, 2012; Desouza and Simons, 2014; Desouza and Smith, 2014; Desouza et al., 2014; Krishnamurthy and Desouza, Forthcoming), I urge us to mindfully consider key issues within each data dimension. Data collected during these projects includes over 45 case studies, over 60 interviews, over 1000 pages of documents and secondary data analysis, and examination of over 70 technology solutions that span mobile apps, online crowdsourcing platforms, sensors, analytical and visualization technologies, and associated urban technologies. The rest of the paper is organized as follows. In the next section, I briefly introduce the three major data trends – big, open, and mobile. Next, I put forth a conceptual framework that will serve as an organizing device to explore how these three data trends impact outcomes when it comes to designing urban spaces. Key issues within four categories—management, technology, infrastructure, and governance—that need to be mindfully accounted for when it comes to the three data movements are presented next. The paper concludes with areas for future research and an outline of practitioner implications.

All Data is Not Equal

All data is not created equally. Each of the three data movements—big, open, and mobile—have had different trajectories and have been shaped by different socio-economic, technological, and global dynamics. I am not a historian, but based on my read of the literature, especially in the urban planning and public administration space, we first had open data. On US President Obama’s first day in office in 2009, he signed a memorandum designed to further open and transparent government efforts. Efforts to provide access to open data have even moved to the international arena. During the 2013 G8 summit, the US, the UK, France, Canada, Germany, Russia, Italy, and Japan committed to work towards making public data “open by default.” The group signed the Open Data Charter to increase data sharing efforts, improve the quality of data, and consolidate efforts to build a data repository (Sinai and Martin, 2013). Next, we had big data. The interest in big data can be traced to the developments in sensor technology that enabled us to collect real-time data at very low cost. The growth in data reservoirs coupled with advancements in data mining and data visualization capabilities enabled organizations to discover latent patterns and gave rise to predictive analytics. Mobile data rose in prominence with the launch of smart phones, opening up of mobile operating systems, and the interest in civic apps (Desouza and Bhagwatwar, 2012; Desouza and Simon, 2014). Since the creation of the iPhone in 2007, mobile “applications” or “apps” have allowed individuals to use their smartphones as a personal computer. Mobile-only users are becoming a strong force in the market. A 2013 Pew Research Center study concluded that 63% of adult cell phone owners use their cell phone to go online and 34% of respondents indicated that they do most of their internet
browsing on their mobile phone. Mobile apps are valued for their convenience and cost efficiency – their use is growing at an astounding rate. Apple’s App Store holds more than 800,000 apps and the Android app store has approximately 750,000 apps (McCracken, 2013).

The three data movements have been fueled by an interest in crowdsourcing (Desouza 2012). One aspect of crowdsourcing, especially when we think of the urban context, is the rise in challenges and competitions that seek to engage the public in the analysis of data towards creating solutions of value to the community. Simply put, challenges are crowdsourced competitions designed to empower citizens to tackle problems or realize opportunities. Challenges use incentives, monetary and non-monetary, to drive participation in the competition. Prizes have been used as incentives to spur achievement and recognize excellence. Prizes assume that individuals and groups are motivated by desires to obtain rewards and avoid sanctions (Dixit, 2002). There is a prominent argument in the literature on incentives suggesting that prize awards can be a mechanism for accelerating technological development to achieve innovation (Kalil, 2006; Kremer, 1998; Wright, 1983; Brunt, Lerner, & Nicholas, 2008; Kay, 2011). Innovative approaches using new technologies can provide citizens a way to directly participate in governance and improve public policy. Challenges are a useful tool for motivating citizen engagement and deliberation. They bring citizens together in a competitive scenario to address a problem, and thereby serve as a catalyst for exploiting collective intelligence. Challenges can be designed to extend far beyond the solution generation phase – citizens are often engaged with the selection and implementation of solutions as well (Desouza, 2012). Bottom-line, through the use of crowdsourcing, we can leverage collective intelligence towards all aspects of data management from the collection of data to its analysis, visualization, and application.

**Big Data**

Urban informatics discussions often focus on big data. Effective use of big data gives us the ability to collect data from our environment in real-time, analyze massive data sets, make real-time decisions, and create policies that operate in real-time. The growth in sensors and other data collection tools has spurred much of this interest, e.g., today we have scanners that can be used for automated tolling of cars and sensors on our phones that can give us context-specific information. Public agencies involved in the planning and administration of cities have traditionally been the repositories of data. This data has traditionally not been evaluated in any meaningful way besides the processing of transactions (e.g. billing purposes) for the simple reason that public agencies have lacked the necessary technical infrastructure and fiscal resources.

The increasing popularity of social networks has added a new dimension to the big data conversation. Today, citizens are able to connect with social networks (e.g. Facebook and Twitter) to share real-time information about their environments. Data shared on social networks range from checking in at restaurants to their travel routes (e.g. through the commonly used Facebook Check In/Place tool). Citizens also frequently share their experiences and reviews of services via social networks. Additionally, online civic platforms are connecting people and solving challenges. These civic engagement tools succeed because they cater to citizens and streamline the interaction process between public agencies and individuals. In 2010, Pew found that 22 percent of all adults receive alerts about local issues and 20 percent of all adults used digital tools to talk to their neighbors about community issues (Smith, 2010). Nearly in one in ten social network users joined an online group that focuses on community issues in the past year. Pew also found that Americans under age 40 were just as likely to give donations to disaster relief through digital means as they are through traditional means like the phone or postal mail (Purcell and Dimock, 2011). These trends illustrate that citizen engagement via online tools is growing and the capacity of cities and communities to meet their unmet needs through use of technology is a fertile ground for research. Similarly, Rio de Janeiro, Brazil put a traffic system in place which coordinates information from more than 20 city departments with technical support from IBM. By using sensors and video feeds around the city, this centralized system produces real-time maps in crisis situations. Utilizing real-time
information, the city government has reduced the response time to crises by about 30 percent. In addition, IBM employs Facebook and Twitter to inform drivers about traffic flows and alternate routes on the occasions of major festivals, matches, and even during accidents (Carter, 2012).

National Association of State CIO’s (NASCIO) (2012) August report notes that some CIOs “have not yet seen an imperative for big data. While the idea of big data is gaining traction, implementation may not be in reach because of staff training or lack of database management tools”. In a response to NASCIO’s annual survey of state Chief Information Officers (CIOs), one CIO explained that “we have some analytical tools, but not of the scale we need to use for mining big data sets”. Furthermore, an analysis of the survey revealed that state governments are experiencing a learning curve and “State CIOs are experiencing a delicate balancing act of maintaining older, but necessary legacy technology while embracing emerging innovative technologies such as big data, cloud computing, mobile devices, social media and public safety broadband” (“Where is big data in state CIOs' strategic planning?”, 2012).

Leading cities have invested in leveraging big data. For instance, New York City (NYC) is employing “risk-assessment and predictive resource allocation for all 60-some city agencies, ranging from fire inspections, building inspections, audits by revenue collecting agencies or business licenses, all the way to entrepreneurial assistance, in the form of identifying locations where there’s a specific combination of businesses, to having a suppressive impact on certain catastrophic outcomes, like crime or fire or water main breaks or things like that”. NYC receives approximately 20,000 to 25,000 complaints every year for something called an ‘illegal conversion’ (where an apartment or a house zoned to accommodate six people for safety reasons is used to accommodate more than 20 or so people). Since NYC’s Department of Buildings only has around 200 inspectors to deal with this issue, a data analytics database was created to compile a ‘greatest catastrophic risk’ list using property tax, building type, age, and other variables. With this new system in place, the Department of Buildings inspection team is increasingly effective, i.e., about 70-80 percent of building investigations involve risk prone conditions compared to 13 percent historically (Howard, 2012).

In many aspects, big data is still in a research and development phase. Discoveries are needed to ensure that we have effective and efficient tools to mine big data. In 2012, the Obama administration announced the Big Data Research and Development Initiative and proposed to invest $200 million in developing big data analytics capabilities within the public sphere. Many federal agencies joined this initiative due to their interest in developing their capabilities to mine valuable information from data within their organizations, transform service delivery, and tackle intractable social challenges. For instance, the Department of Defense announced that it would spend $25 million per year for four years to develop a XDATA program with computational tools and techniques to mine massive data in different formats (structured and semi-structured) within military departments (U.S. Office of Science and Technology Policy, 2012).

Open Data
Simply put, open data programs center on making data available to the public. While the impetus and goals of the programs might vary slightly, they mostly center on increasing transparency and collaboration between public agencies and citizens. In addition to promoting transparency and fostering collaboration, public agencies have begun to realize that open data efforts generate economic value. For instance, by publishing bidding information online, the Texas State Administration saved more $15.2 million in prison food costs by renegotiating its contract (US PIRG Education Fund, 2012). Through the creation of apps based on 'open' data, citizens are able to create business enterprises and solutions designed to increase the efficiency and effectiveness of public services, e.g., through apps that increase public transit ridership by providing real-time bus tracking data. Consider three examples:
In 2010, a massive earthquake hit the city of Christchurch, New Zealand. The 7.1 magnitude earthquake destroyed the city’s critical infrastructure, i.e., buildings, power supply, cables, etc. While city officials, businesses, and residents were keen to begin the recovery processes, they faced a grave challenge: lack of data. City officials quickly realized that they did not have a comprehensive data inventory of its critical infrastructure to begin their urban planning. Lack of data crippled the recovery process; in 2013 the New Zealand government initiated the Canterbury Spatial Data Infrastructure Program (SDI). This initiative aims to facilitate collaboration between public and business sectors by encouraging them to share location-based data (Land Information New Zealand, n.d.).

The NYC BigApps Competition offers a grand prize of $50,000 and other prices to individuals, groups, private companies, and non-governmental organizations that develop apps using the NYC open data to improve the functioning of the city. The goal of this initiative is twofold. First, the initiative aims to provide improved access to government information and improve transparency. Second, the BigApps competition is designed to encourage innovative ideas and generation of intellectual property from sources external to government, i.e., with the involvement of public, startups, private companies, and non-profit organizations. Participants who register on the BigApps website can submit ideas, vote on submissions, and join discussion boards. Recently, NYC announced its BigApps 3.0 winners. The BigApps 3.0 initiative is built on the previous two BigApps competitions that encouraged software developers to develop apps “for the web, a personal computer, a mobile device, SMS, or any software platform broadly available to the public” (“NYC BigApps 3.0”, 2011).

In recent years, international agencies have begun to reach out to the public to develop apps aimed at solving global issues such as poverty, political riots, and economic growth. For instance, in April 2011, the World Bank announced its Apps for Development Competition (“Apps for Development”, n.d.). This competition encouraged software developers and practitioners to develop apps to visualize and create maps using various datasets, e.g., World Development Indicators, Africa Development Indicators, and Millennium Development Goal Indicators. Since all this data was freely available through World Bank open data initiative web portal, developers were able easily access requisite information and to submit proposals ranging from web applications to mobile apps. Once the idea is submitted, World Bank screens the submission and posts it on its website for public voting. The most popular app wins ‘Popular Choice Award’. Submissions were judged based on their usefulness in achieving the Millennium Development Goals (MDG).

While there are clear benefits to opening up data reservoirs, we must remember that making data open is no easy feat to accomplish. Agencies must invest resources to make data open. First, most of the public agencies’ data were collected for internal decision-making purposes not for public release (Dawes and Helbig 2010). Second, different agencies use different strategies and methods to collect data. As a result, users should have a clear contextual understanding of how data were collected before they use the data (Dawes and Helbig 2010). Third, there are considerable costs associated with converting the closed data into a machine-readable format (Dawes 2010; Noveck 2012). Fourth, metadata should be updated in a timely fashion for relevance, and users should be informed of any changes in the datasets (Daconta 2009; Robinson, Yu, and Felten 2010). Fifth, care must be taken to ensure that the data being released adheres to privacy and security requirements, which requires considerable investment of time and energy (Lewis and Perricos 2013). Given the increased pressure on public agencies to do more with less, agencies are struggling to meet their core mission objectives, and there are limited, if any, slack resources. As witnessed during the most recent shutdown of the US federal government in October 2013, all open data platforms, including Data.gov, were considered non-essential services and were shutdown. Moreover, simply making data open is not sufficient (Daconta 2009; Roberts 2012; Robinson et al., 2010). Public
agencies need to invest in raising awareness of open data programs and equipping citizens with the capacity to leverage the data (Noveck 2012; T. Roberts 2012).

Liberating data can promote open innovation (Mergel and Desouza, 2013), which involves public participation in policy making and collaboration between public agencies and citizens to address complex social problems. Thus, in order for public agencies’ open-data initiatives to generate economic and social value, agencies should encourage active civic participation so that citizens use the data for their needs and those of others in society. Co-production models are necessary so that the public feels a sense of duty and responsibility to participate in the policy processes (Clark et al. 2013; Meijer, 2011). The co-production approach strongly advocates that users should be engaged in the planning processes to improve service delivery (Bovaird 2007; Clark et al. 2013; Whitaker 1980). For instance, as of October 2013, the Chicago Data Portal (https://data.cityofchicago.org/) provides access to 979 datasets from various city government agencies. The Chicago government released crime data dating back to 2001. In 2010, the Chicago Tribune News Apps Team utilized this data to develop the Chicago Tribune’s Crime app (http://crime.chicagotribune.com/) to inform public about the crime rate in the neighborhood, community, and public transit. Since 2010, the Chicago Councilmatic app (http://chicagocouncilmatic.org/) built by Derek Eder and Forest Gregg at Open City has allowed citizens to browse, comment on, and subscribe to Chicago city council initiatives. Similarly, as of October 2013, the San Francisco Data Portal (https://data.sfgov.org/) provides access to 721 datasets. In 2012, Appallicious, a startup company, created SFRECPARK for the city and county of San Francisco. This app allows users to search for parks and recreation facilities located in the city of San Francisco. The app allows users to make picnic table reservations, volunteer, and donate. Using this app, users can also purchase ticket and multi-day passes to classes, museums, and tours.

**Mobile Data**

Mobile data gives us the most real-time and updated information on citizens that traverse a city. For instance, the Indian government introduced SMS services allowing people to make reports concerning excessive emissions from busses and other vehicles (OECD Report on M-Government, 2011). Advancement in mobile technologies (e.g., 3G and 4G networks) has empowered people to generate information on the go. Citizens have become sensors with an ability to provide real-time information on a multitude of issues. Consider a few statistics:

- Global smartphone users exceeded 1 billion in 2012 (Reisinger, 2012) and were expected to grow 40% in 2013 to 1.4 billion (ABIresearch, 2013).
- In 2013 tablet users reached more than 250 million, with an annual growth rate of 125% (ABIresearch, 2013).
- According to Mary Meeker of Morgan Stanley, “[by 2018] more users will connect to the Internet over mobile devices than desktop PCs” (Ingram, 2010).
- Apple tablet and smartphone users have downloaded over 50 billion apps, and Android tablet and smartphone users have downloaded close to 50 billion apps (Dediu, 2013). This equates to almost 14 apps downloaded for every person on the planet (U.S. Census Bureau, 2014).

Mobile data has spurred the development of mobile ‘apps.’ By default most public agencies think about how to connect with citizens through mobile apps (Desouza and Bhagwatwar, 2012). Mobile apps cover the entire spectrum of urban life from transportation to crime alerts and parking locators to parks and recreation. For example, Google and Microsoft mashed up health data from the United States Department of Health and Human Services with geo-spatial data to create search engines to help parents searching for information about hospitals to people looking for health related information (Noveck, 2012). While we can expect the number of apps to increase in the future, what is more interesting is the intensifying dependency on apps as the gateway to information on an urban space. Moreover, mobile technologies offer communication channels where wireless and Internet is not a viable option (OECD Report on M-Government, 2011). During the aftermath of the Haiti earthquake, the Ushahidi team provided an
international SMS number to connect the Haitian diaspora and coordinate relief efforts. They even published online maps showing damage to buildings and connected these images to satellite images showing open roads to coordinate relief efforts. Several individuals added to these maps, providing real-time information to help public agencies, NGOs, and international aid organizations coordinate their efforts toward earthquake victims (Forrest, 2010). Similarly, a group of researchers from Columbia University and Karolinska Institute (Sweden) analyzed mobile data to predict refugee movement and associated health risks after the 2010 Haiti earthquake. The researchers predicted the destinations of more than 600,000 people, which helped public agencies and international aid agencies to better coordinate their efforts and address the needs of the people (World Economic Forum, 2012).

When considering urban informatics, mobile data gives us the ability to get location-specific and activity-specific data from agents (e.g. citizens) and objects (e.g. phones/sensors). Peer-to-peer or device-to-device exchange of data allows for real-time collaboration and coordination within cities. Citizens can participate in the creation of mobile data through efforts such as community mapping.

**Conceptual Framework**

The conceptual framework is captured in a figure 1. First, we have the three kinds of data: big, open, and mobile. Each type of data has unique features and provides us with varied insights. For example, when people have access to data, they can create novel solutions to traditional urban challenges such as crime, poverty, transportation, and etc. In the city of Memphis, researchers from University of Memphis approached the local police department and asked them for historic crime data. These researchers cross-tabulated crime data with geospatial data and created crime maps for the city. With this new map in place, the city recorded a drop in crime by 30 percent in 2012 compared to 2006 (Badger, 2012).

Next, we have the outcomes that we seek when we think of designing the next-generation of cities. Most of us would agree that we should have cities that are *smart*, *livable*, *sustainable*, and *resilient*. We can conduct a deep dive into each of these desired outcomes. A city that is *smart* will enable its citizens to traverse its environments in the most efficient and effective manner as they conduct their activities. Similarly, a city that is *livable* will use data to ensure that its citizens and organizations have an adequate standard of living, access to critical resources and services, etc. Cities that are *resilient* will use data and information to ensure that they can sense impending disasters, both immediate and long-term, and in the cases where these events cannot be avoided, appropriate contingency plans can be put in place. Finally, *sustainability* calls for ensuring that a city is responsible in terms of how its resources are consumed, conserved, replenished, and managed to ensure that the city does not compromise its future.
In the middle of Figure 1, we have the key design interventions—management, technology, infrastructure, and governance—that one needs to carefully consider if they are to leverage data towards the goal of building cities that are smart, livable, resilient, and sustainable. It is not possible to cover all permutations of combinations when we consider the various components of the conceptual framework in this paper. I will focus my discussion on some of the most salient issues that need to be contemplated as we think of urban informatics.

**Management**
The ability of a public agency to conduct *data management* effectively and efficiently towards achieving the desired outcomes for a city can play a critical role when it comes to leveraging the three types of data. Data management encompasses all aspects of data creation, storage, transfer, curation, analysis, visualization, implementation, and even destruction (and/or archiving). Public agencies face an uphill battle when it comes to data management (Desouza, 2014a). Having interviewed Chief Innovation Officers (CIOs) and Directors of IT departments across federal, state, and local levels in the US, I found that while public agencies have realized the value of investing in big data projects, they are facing several challenges including developing data governance, attracting data scientists, and managing issues related privacy and security.

Public agencies collect massive volumes of data. However, much of this data is collected for internal organizational purposes only and without the intention of sharing the data with other agencies and the general public. Hence, to move towards publishing organizational data, agencies have to think critically about their data collection, sharing, and organizing strategies. Another critical concern: converting historic data buried within the organization (which generally requires large investments). Even if agencies are willing to invest resources to convert their data reservoirs, skepticism remains about if these data will offer any new insights. Much of the information was collected for a specific purpose; maybe the records do not have complete contextual information to help analysts run meaningful analyses.

Public agencies have struggled to modernize their data management programs even when they received funding to do so (e.g. the FBI Virtual Case File and the IRS Business Modernization Project). What is alarming here is the fact that even when contending with structured and tightly controlled data, agencies have fared poorly when it comes to data modernization efforts. One can only speculate how dismal results will be when we consider trying to manage data that is unstructured, of high variety and volume, and from sources that are beyond an agency’s control. Despite my skepticism, there are a few success stories. For instance, in 2009, the New York City recorded 158 deaths due to drug overdoses involving opioids. City administrators analyzed large records of Medicare data and found that the use of opioid drugs increased over tenfold over the past 20 years. Furthermore, one percent of pharmacies received 25 percent of opioid Medicaid reimbursements. This information helped health and law enforcement departments to manage opioid drug usage, thereby preventing deaths due to opioid drugs (Saul, 2011).

Another concern plaguing the collection of data from social networking sites is the lack of access to complete information. Consider the simple example of liking/sharing a post on Facebook. In less than a second, you can like something and unlike it or share a post and delete it. This is a critical concern for researchers who may lack an ability to see how things have changed. Collecting data from these sources requires more than simply merging data from different sources. In addition, private companies which collect and sell data for premium may develop collaboration with these social network sites to limit the open access to the data generated through these platforms. A recent congressional bipartisan ‘Privacy Caucus’ headed by representatives Ed Markey and Joe Barton is developing strategies and governing principles to ensure transparency and the creation of physical world values to accompany the governance of the digital world (Singer, 2012). As a first move towards this end, the committee initiated inquiry into self-regulated companies, which collect, analyze, buy, and sell personal data (Petulla, 2012). Despite the
efforts by many public agencies to design strategies and policies to govern access to digital information, many of the initiatives are in early stages of inception with no clear success criterion to evaluate them.

**Technology**

The technology component deals with all aspects of technical infrastructure within a city and the agencies that are responsible for designing, planning, and managing a city. Unchecked overreliance on mobile technology-driven policy responses can potentially widen the gap between segments of the population that can afford these emerging technologies and those who cannot. Consider this case: the city of Boston introduced “Street Bump” app, which automatically detects potholes and sends reports it to city administrators. As residents drive, this app collects data on smoothness of ride, which could potentially aid in planning investments to fix roads in the city. However, after the launch of the app, it was found that the program directed crews to wealthy neighborhoods because people were more likely to have access to smartphones (Rampton, 2014). This example illustrates that public agencies cannot simply leverage technologies to address urban challenges; they need to think through several dimensions. Who are the users of the application? Is it representative of all sections of society? Also, data ownership becomes a crucial issue. Who owns the data? How can people be used as sensors without violating their privacy? Incidents like these will affect cities’ resilience and livability.

Technology can also be easily manipulated. For example, in 2011, the Obama administration proposed the Keystone XL pipeline project to carry tar sands oil from Alberta down to Texas (Olson and Efstathiou Jr., 2011; Sheppard, 2011). To counter the expressed concerns among stakeholders, the American Petroleum Institute and other lobbyists were able to manipulate social media sentiments to show support for the project. By using ‘fake’ Twitter accounts, they were able to send an inordinate number of tweets to show support for the project, which did not accurately represent the sentiment on the ground. While this example is perhaps more nefarious than typical, it does reflect a real-concern regarding the potential for manipulating data to sway public policy.

In order to analyze data of this scale, researchers should apply methodological rigor grounded in sound theories and be transparent about the process of data analysis (Boyd & Crawford, 2011; Petrkeil, 2013). Thus, analysis of big data requires new kinds of significance tests and other validation techniques that gauge the temporal variability to discover patterns and relationships. As noted earlier, a large volume of data is generated through social network sites. Networks of people discuss, express their views, share/like posts, and view information online. But these networks are not representative samples of global population. Consider the case of Twitter or Facebook. A person can have multiple accounts or multiple people can use one account. First, uncertainty about account usage and who engages in these platforms makes it difficult to use the data for interpreting claims about human behavior (Boyd and Crawford, 2011). Second, while collecting data from these sites, researchers are required to make subjective judgment about how to code a large volume of unstructured data.

Big data movements have created debates about the value of traditional research design. Some have argued that as analysts mine large volumes of data, simple correlations could reveal several useful insights for decision making. However, this is not the case. Consider this example: in 2010, data scientists at Google revealed that by crunching large volumes of data-related search terms, they were able to detect patterns of swine flu. More importantly, they accurately estimated the actual number of people who contracted swine flu. This report surfaced weeks before the official announcement from the US Centers for Disease Control and Prevention (CDC), and caused many to question the CDC’s method of disease prediction (Loukides, 2010). Further, a recent study has revealed that the Google team has been overestimating flu outbreaks since 2011 – their prediction is two points off compared to the CDC estimates. Additionally, by employing its traditional methods of estimation, the CDC has been accurately predicting flu outbreaks (Lazer, Kennedy, King, and Vespignani, 2014).
This example highlights several issues with big data analytics. While big data analytics has the potential to predict outcomes, we need to develop new methods of data analysis. Traditional methods, which involve building theories, testing hypotheses, and interpreting results may be ill-fitted when it comes to big data analytics. Researchers need to cultivate new ways of developing theories and running analysis to detect patterns. Moreover, as with deriving any meaningful analysis, detecting patterns will depend upon the ability of researchers’ to separate noise from data, run multiple analyses, conduct robustness checks, and validate results. Also, we need to critically and evaluate the potential of big data tools in reducing costs, optimizing operations, detecting patterns, and predicting outcomes. Care must be taken to ensure methodical rigor and disclose data analysis processes when applying big data solutions to complex problems.

Protecting personally identifiable information about citizens is a fundamental responsibility of public agencies (Dunleavy, 2012). But until now, much of the initiative towards big data collection, analysis, and storage has been driven by private sector. Moreover, public agencies lack full knowledge about what type of information is collected by these private entities (Manovich, 2012; Stanley, 2012). Private companies such as Netflix, Walmart, Amazon, and Target collect large amount of information about customers to tap into behaviors of these individuals to provide tailored services. Target, for instance, assigns an identifier to each customer, which is then mapped to their credit card number, contact information, and their purchasing history at any Target store (Duhigg, 2012). Using this repository of information, Target analysts can predict a range of information about the customer, e.g., pregnancy and other health-related information. The discovery about customer’s needs is used to offer personalized ads to each customer. In a shocking incident recently, Target accurately figured out about a teen’s pregnancy even before her father came to know about the issue (Hills, 2012).

Many organizations limit the data that they share. Companies like Facebook, Twitter, Flickr, Amazon, and Google do not allow full access to their Application Programming Interface (API – which is a set of commands used to retrieve information from databases). Researchers can only get access to a small percentage of the information (Manovich, 2011). For example, in 2007, Facebook launched Beacon, an advertisement system which allowed access to external websites to post information to user’s news feed about their purchases online (Boyd, 2010). Consider this simple case: if you are planning a trip to Boston and purchase your ticket, the Beacon will automatically post this information on your news feed revealing your travel plans to your friends. Unfortunately, many people could not figure out how this was happening, and worse, many people were not even aware of this function. This invariably resulted in a drastic shift in online behaviors (Boyd, 2010). In a 2008 lawsuit against Facebook and its partners, it was noted that a man who bought a diamond ring to propose his girlfriend realized that the information about his recent purchase was posted on his account before he got a chance to propose his girlfriend. This feature shattered the man’s dream, and caused considerable outcry from online privacy advocates. Despite the various lawsuits against Facebook and its partners, Facebook continues to collect massive volume of data.

Similar dynamics play on many online sites. For instance, millions of users utilize Google search engine every day to find trivial information (e.g. word meanings) as well as critical information (for e.g. awareness about particular diseases). During the flu seasons, many people search for flu related information. Google harnessed search data to observe patterns during the Swine Flu epidemic. They predicted trends in Swine Flu even before the Centre for Disease Control (Loukides, 2010). Google could derive patterns by discovering a relationship between numbers of people who search for flu related topics to actual numbers of people who have contracted the flu. Similarly, the amount of information collected through Facebook could help researchers to predict and evaluate social issues ranging from predicting local criminal network to global terrorist networks (Howard, P., 2012). Even when the public files motions to obtain full access to companies’ APIs, a lack of global regulatory standards relating to privacy laws makes public disclosure of data difficult (Boyd, 2010; Howard, P., 2012). Given the lack of
standards and norms governing access to information, it is difficult to ascertain the reliability of data obtained through these sources. Or worse, fear the chances of information being tainted. Public agencies have inducted several acts and laws to govern issues of privacy. However, in the wake of unprecedented data generated through global connection, these laws often fall short in fulfilling their promise. Access to this volume of information offers unprecedented power to the people who own these large datasets. This can create global dependencies where few big players will have the capacity and resources to ‘rule the world’.

In the context of urban informatics, one needs to ensure that data collected from citizens, especially data collected on behavior of a sensitive nature (e.g. through the use of video cameras on streets) is held securely and used for its intended purposes only. A recent study by MIT’s Whitehead Institute noted that by cross-linking chromosome public data, they were able to identify 50 participants who submitted their DNA (Rijmenam, 2013). The big policy lesson is that we need to reexamine our approach to de-identify personal information from public datasets. This calls for developing sophisticated data management approaches and investing resources in ICTs to mask personal information. Desouza and Simon (2014) detail several risks posed to individuals and organizations that use mobile apps, especially when operating in highly-regulated industries and highly sensitive information. The average citizen does not understand the nuances associated with mobile technology from a privacy and security stance.

Infrastructure
Two key issues play out when we think of infrastructure. The first issue relates to the challenges associated with retrofitting existing (traditional) infrastructure with real-time sensors and data collection technologies. The second issue is with regards to obtaining funding for infrastructure investments within cities (Desouza et al., 2014). The former concern is one of determining graceful ways to transition or renew decaying infrastructure in a sustainable manner. The latter concern is one of securing much needed financial resources for upgrades. Neither concern is trivial.

The UK government invested more than £100 million to provide broadband funding for small businesses in 22 cities (Department for Culture, Media & Sport, 2013). In 2013, NYC Mayor Michael Bloomberg announced several initiatives to expand wireless and broadband connectivity in the city of New York, including the launch of free Public Wi-Fi Networks in ten commercial districts in all five boroughs. Additionally, the Mayor launched WiredNYC, a rating platform to evaluate office buildings’ broadband connectivity. More than 150 buildings signed up for the program (Vorgna, Wood, and Muncie, 2013). The city’s Economic Development Corporation (NYEDC) also ran the Wireless Corridor Challenge to increase wireless connectivity in commercial districts and announced up to $300,000 in funding (NYEDC, 2013). The city’s investment will be leveraged by private-sector commitments totaling over $3.4 million dollars, with the city providing a total of $900,000 in additional funding to assist with the implementation (Vorgna, Wood, and Muncie, 2013).

Van Der Lan, the Director of Amsterdam’s Economic Board, a non-profit organization, found that installing smart grids would cost $20 million less if private and public agencies shared their upcoming investment plans. As he noted, “We have the plans of these different companies. We know what they plan to do and how much money they will put in to realize these plans… When we put all the budgets together we need, say, 100 million [euros]. But when we do it together, it costs us 80 million” (Balch, 2013). Furthermore, as cities do not exist in vacuums, they should reach out to neighboring cities for strengthening their infrastructures. For instance, the city of Helsinki in Finland collaborated with its three neighboring cities to launch the Helsinki Region Infoshare. These cities share their urban data through this portal. Consolidation of urban data across cities provides a fuller picture and helps adopt holistic approaches to urban challenges. Additionally, these cities are making efforts to publish data in a machine-readable format for easy access (Sulopuisto, 2014). As cities are becoming increasingly connected, one agency or one sector cannot effectively solve problems. Furthermore, when cities develop interconnected
infrastructures, they can become more vulnerable. For example, one event could disrupt functioning of the city. Thus, cities should continually assess health of these infrastructures. Cross-agency and sector collaboration will help cities improve its resilience.

Desouza et al. (2014) created a handbook for public managers to understand various financial tools to underwrite investment in infrastructure and smart city projects. Currently, there is a serious deficiency when it comes to financial intelligence when it comes to evaluating contracts for infrastructure management. Consider the now infamous case of Chicago leasing its parking meters to a private concessionaire. Not only was this a bad financial deal from the standpoint of managing public goods but it was made worse by a lawsuit and the ensuing settlement. Existing technologies are in place to help cities manage big data or even design smarter environments through sensor networks, the key issue that remains a challenge is how to fund these investments in a smart and sustainable manner.

Infrastructure investment decisions are even more complicated in fragile cities of the world (Desouza 2014b) where the existing population lacks basic services and the current environment is in decay. As an example consider a place like Mumbai where large portions of the population lack clean drinking water, sanitation facilities, etc. While I do acknowledge that all cities face resource constraints, New York or London face far less constraints than Islamabad or Dhaka. Today, organizations that claim to develop solutions for smart cities, are, in my opinion, out of touch with reality when we consider fragile environments. Cities in the developing world cannot afford the technologies they sell nor have infrastructure at a level of maturity where the solutions make sense. Cities lack basic infrastructure and resources necessary for basic survival needs, and cannot invest the time, energy, or resources to purchase and implement glamorous technological solutions being sold. These innovations are not what these cities need. These cities need a capacity to develop innovations using their own internal resources and capabilities in a manner that is sustainable given their economic, political, and infrastructure realities. Cities need to promote frugal engineering that leverages local knowledge towards creating data-centric solutions (Desouza, 2014b).

**Governance**

Governance includes all aspects of how the three data movements impact how we design and implement policies to manage public goods and deliver public services within cities. Consider this case: after the Sandy Hook shooting incident in October 2011, a local newspaper agency filed an application under the Freedom of Information Act to obtain information about gun owners living in Westchester, Rockland, and Putnam counties in New York. The newspaper agency published an interactive map showing gun owners living in these areas to provide information about gun possession (Johnson, 2011). However, the newspaper agency was met with severe criticism for breaching individuals’ privacy. As an emotional response, some residents published interactive maps with names and addresses of the employees working at the newspaper agency. More importantly, New York public agencies came under critical scrutiny for not protecting residents’ information. This incident highlighted that cities need to (1) develop rules and regulations, (2) educate people about the ethical use of open data, and (3) develop standards and guidelines on proper usage of data.

Big data analytics has the potential to create an unequal society, i.e., they can adversely harm a subsection of a population. For example, a recent report highlighted that in the city of New York, a section of residents were finding it difficult to rent homes. Tenant screening companies obtained data on residents who were sued in housing courts and shared this information with landlords and housing management companies. Landowners and housing companies increasingly used this information to evaluate tenants’ housing applications (Iverac, 2014). While this information greatly helped landlords and housing companies screen applications, it adversely impacted tenants. Boyd and Crawford (2012) note that money and resources are needed to access big data. Data ownership will play a critical role, and individuals who own data will have more power over others. It is clear that access to information empowers people, but
without right control it can adversely affect cities’ livability. Thus, cities need to think creatively about balancing access to information and protecting privacy. Consider this case: residents are increasingly seeking the advice of private firms that offer one-on-one advice when it comes buying homes. Individuals have reported that these firms provide accurate information about localities based on their customer preferences. While this is good news for people searching for homes, this raises challenges for cities. Residents have a fundamental right to information; however, some information may result in creating situations where residents are likely to choose “people like us” (Prevost, 2014). This affects a city’s livability and resilience.

Concluding Thoughts
This paper has outlined a conceptual framework for analyzing key issues when it comes to leveraging to big, open, and mobile data in the context of urban informatics. The current scholarship on urban informatics has largely been dominated by work that focuses on how to build tools and technologies to leverage real-time data towards making devices, organizations, and individuals smarter when it comes to their activities and movements in urban spaces. While good, the technocrat view should be complemented with a thoughtful consideration of governance and socio-economic issues.

This paper has raised a number of issues that require further investigation and practitioner attention. First, we need to get more sophisticated in our understanding of unintended consequences big data analytics can cause from a socio-economic stance. Will big data analytics help us build more livable cities for the masses or will it just deepen existing segregation in our communities? Second, when it comes to open data, the key question remains how does one measure the economic impact of these programs? Making data open is a costly undertaking and public agencies already are operating in resources-strained environments. Today, open data programs are fashionable and many cities are just ‘Keeping up with the Joneses’ when it comes to making data available. Third, mobile data allows us to send and receive real-time data to citizens. Today, the mobile app space is crowded and this trend is expected to continue. Research is needed to understand the dynamics of the apps marketplace, the impact of apps on changing behaviors in urban spaces, etc. For e.g. more apps require control of an individual’s device and access to sensitive data (e.g. phone numbers, etc.), research has yet to examine the security and privacy implications that arises from this kind of data provision.

The field of urban informatics is going to flourish as cities expand and grow. New forms of data and technologies are giving us new insights towards the planning and governance of cities. Ultimately, however, our success in leveraging data in the urban context depends on our ability to carefully consider key social and governance dilemmas.

References


Dunleavy, M. (2012). Big data: The public sector is still lagging behind. Available at: http://www.publicservice.co.uk/feature_story.asp?id=21309


Howard, P. (2012). Let’s nationalize Facebook: Only then will the social network protect users’ rights and share valuable data with researchers. *Slate*. Available at: http://www.slate.com/articles/technology/future_tense/2012/08/facebook_should_be_nationalized_to_protect_user_rights_.html


Kearney, 2013. The mobile economy 2013. Available at: 
http://www.gsmamobileeconomy.com/GSMA%20Mobile%20Economy%202013.pdf
Lewis and Perricos 2013
Loukides, M. (2010). What is data science? O’Reilly Radar. Available at: 
Marco. (2011, January 11). Open data vs optimistic assumptions about citizens data literacy. Available at: 
McCracken, H. (2013, April 16). Who’s winning, iOS or Android? All the numbers, all in one place. *Time*. Available at: 
Novek, B. (2011, April 7) What’s in a Name? Open Gov and Good Gov. The Huffington Post. Available at: 
http://crookedtimber.com/2012/07/05/open-data-the-democratic-imperative/
Obama, B. (2009, January 21). Memorandum on transparency and open government. Available at: 
http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment
Olson, B., and Jim Elstathiou Jr. (2011, November 16). Enbridge’s pipeline threatens trans Canada’s Keystone XL Plan. *Bloombrg Bussiness Week*. Available at: 


Digital Infomediaries and Civic Hacking in Emerging Urban Data Initiatives

Piyushimita Thakuriah (Vonu)
Director of UK ESRC Urban Big Data Centre, Halcrow Chair of Transport, Professor of Urban Studies and Associated Professor of School of Engineering
University of Glasgow, UK
Piyushimita.Thakuriah@glasgow.ac.uk

Lise Dirks
Senior Associate, Urban Transportation Center
University of Illinois at Chicago
ldirks1@uic.edu

Yaye Keita-Mallon
PhD Candidate, Department of Urban Planning and Policy
University of Illinois at Chicago
ykeita2@uic.edu

To be published in Proceedings of NSF-sponsored workshop on Big Data and Urban Informatics held in Chicago, Aug 11-12, 2014
Abstract

This paper assesses non-traditional urban digital infomediaries who are pushing the agenda of urban Big Data and Open Data. Our analysis identified a mix of private, public, non-profit and informal infomediaries, ranging from very large organizations to independent developers. Using a mixed-methods approach, we identified four major groups of organizations within this dynamic and diverse sector: general-purpose ICT providers, urban information service providers, open and civic data infomediaries, and independent and open source developers. A total of nine organizational types are identified within these four groups.

We align these nine organizational types along five dimensions accounts for their mission and major interests, products and services, as well activities they undertake: techno-managerial, scientific, business and commercial, urban engagement, and openness and transparency. We discuss urban ICT entrepreneurs, and the role of informal networks involving independent developers, data scientists and civic hackers in a domain that historically involved professionals in the urban planning and public management domains.

Additionally, we examine convergence in the sector by analyzing overlaps in their activities, as determined by a text mining exercise of organizational webpages. We also consider increasing similarities in products and services offered by the infomediaries, while highlighting ideological tensions that might arise given the overall complexity of the sector, and differences in the backgrounds and end-goals of the participants involved. There is much room for creation of knowledge and value networks in the urban data sector and for improved cross-fertilization among bodies of knowledge.

Keywords

Big data, open data, urban digital infomediaries, industry convergence, urban ICT entrepreneurs, civic hacking, smart cities, location based services, location based social networks, data scientists, application developers, open source
1. Introduction

There has been a surge of interest recently in urban data, both “Big Data” in the sense of large volumes of data from highly diverse sources, as well as “Open Data”, or data that are being released by government agencies as a part of Open Government initiatives. These sources of data, together with analysis methods that aim to extract knowledge from the data, have attracted significant interest in policy and business communities. While public and private organizations involved in planning and service delivery in the urban sectors have historically been users of urban data, recent developments outlined below have opened up opportunities for policy and planning reform in public agencies and for business innovation by private entities. Such opportunities have also attracted innovative new organizations and facilitated new modes of ICT entrepreneurship. The objective of this paper is to examine the diverse organizations and networks around such emerging sources of urban “Big Data” and “Open Data”.

While there are many explanations of Big Data, it is the term being applied to very large volumes of data which are difficult to handle using traditional data management and analysis methods (Thakuriah and Geers, 2013; Batty, 2013), and which can be differentiated from other data in terms of its “volume, velocity and variety” (Beyer and Laney, 2012). It has also stimulated an emphasis on data-driven decision making based on analytics and data science (Provost and Fawcett, 2013) that has the potential to add to the value brought about by traditional urban and regional modeling approaches.

Urban Big Data can be generated from several sources such as sensors in the transportation, utility, health, energy, water, waste and environmental management infrastructure, and the Machine-to-Machine (M2M) communications thereby generated. The increasing use of social media, using Web 2.0 technologies, personal mobile devices and other ways to connect and share information has added to the vast amounts of socially-generated user-generated content on cities. Open Data initiatives adopted by city governments are leading to an increasing availability of administrative and other governmental “open data” from urban management and monitoring processes in a wide variety of urban sectors. Privately-held business transactions and opinion-monitoring systems (for example, real estate, food, or durable goods transactions data, data on household energy or water consumption, or customer reviews and opinions) can yield significant insights on urban patterns and dynamics.

The increasing availability of such data has generated new modes of enquiry on cities, and has raised awareness regarding a data-driven approach for planning and decision-making in the public, private and non-profit sectors. Urban Big Data has stimulated new data entrepreneurship strategies by business, independent developers, civic technologists, civic hackers, urban data scientists and the like, who are using data for civic activism, citizen science and smart urban management in novel new ways. However, very little has been written about ways in which such data-centric developments are being organized and delivered, particularly the organizations and networks involved.

In this paper, we qualitatively review organizations in the emerging urban data sector, with the purpose of understanding their involvement in production and service delivery using the data. We use the term “urban digital infomediaries” to describe those enabling organizations and networks which are fostering ICT-based data-centric approaches to studying, managing and engaging in cities. Many organizational modes are prevalent in this landscape, including well-established multinational ICT-focused businesses as well as informal networks of individuals or freelance ICT developers with interests in data and cities. Hence the traditional definition of an infomediary as an “organization” may need to be expanded to include these unstructured and informal activities.

Given the dynamic growth that is being experienced in the urban data sector, several new cross-cutting technology solutions and service delivery processes are emerging. These solutions are being adopted by a wide spectrum of infomediaries with vastly different mission and customer/client focus, leading to similarity in products and services, skills of the workforce involved, and to transcending and blurring of traditional functional lines and organizational/market boundaries. More and more ICT
products use similar intermediate inputs over time and meet similar demands, greatly facilitating such convergence (Xing, et al, 2011; Stieglitz, 2003). Convergence in ICT companies is a well-established area of research, where authors have used conceptual frameworks and case studies to analyze the mechanism of the convergence and its implications. The analysis undertaken here is partly driven by this strand of literature, and we attempt to understand where traditional organizational boundaries are blurring as new opportunities for collaboration arise in the urban data sector.

The paper is motivated by a need to understand city-centric ICT services in order to discern trends relevant for governance, business development and service provision. We are also motivated by convergence in processes for potential partnership building and alliances in providing urban services. The analysis is composed of two major components. First, we undertake a broad, qualitative assessment of organizations and their formal and informal activities regarding urban Big Data. Second, we use a mixed-methods approach to take an in-depth look at a sample of organizations to understand organizational mission and stated objectives, major interests and functional activities, skills and interests of the workforce involved, and services and products. The end result is a description of organizations in this emerging sector along multiple dimensions, as opposed to the traditional categorization such as industrial classification systems.

The paper is organized as follows. In Section 2, we describe our research approach, followed by the analysis of urban infomediaries in Section 3. Conclusions are given in Section 4.

2. Analysis of Organizations: Research Approach

Our objective is to examine organizations that are involved in urban data and the myriad activities that relate to the urban data infrastructure. A mixed-methods approach is utilized, consisting of qualitatively examining the literature and using personal experience and by communicating with experts, as well as quantitatively analyzing material from websites of selected organizations using text mining. The qualitative assessment helped to identify the major groups of stakeholders in the urban data landscape, types of products or services generated, skillsets of professionals involved and evolving professional networks.

In order to understand the work of the urban digital infomediaries in greater detail, a database of webpages of 139 public, private and non-profit ICT-focused organizations and informal ICT entities was constructed. The webpages were collected using snowballing techniques, starting with a list of organizations involved in city-related activities known to the authors. Additional organizations were identified through Internet search using keywords such as “open data”, “smart cities”, “big data”, “civic”, “open source”, “advocacy” and so on, as well as with keywords relating to modes such as “participatory sensing”, “crowdsourcing”, “civic engagement”, “public engagement” and related terms. Pages within websites retrieved for this purpose include, among others: (1) about us, mission statement, products or services, or similar page(s) of organizations which describes the organization; and (2) terms of service, privacy policy or related pages which describe the organization’s terms or policies regarding information use and data sharing.

The first step was to manually label and categorize the type of organizations, sector (public, private, non-profit), major functional interest or domain area and types of services offered, as well as policies and markets. It also allowed us to make an assessment of the skills sets of the urban-centric workforce involved in the organizations, although this aspect was informed by additional reviews and judgment. This led to the identification of four major groups of infomediaries, which were then organized into nine subgroups based on stated missions and interests by organizations.

One use of the database was to understand, using text mining, the emphasis of the organizations with regard to their activities and processes that may not be apparent from their stated mission and objectives. For example, an ICT-focused organization may indicate that it is in the business of “smart
it is possible that it is involved in the smart cities agenda by helping to empower residents to connect to city governments or it may be focused to a greater extent on building the technologies to make such empowerment technically possible. The assumption is that specific focus in the work of “serving communities through smart city technology” can be discerned from the text contained in the websites of organizations.

A statistical clustering approach was applied to the data retrieved from the webpages. This process involved the following steps: (1) creation of a database of words extracted from the webpages of organizations after standard preprocessing to convert text to data (i.e., where each word in every document becomes a column of a rectangular database, with the rows being the organization ID, and with the cells of the final database giving a 1 to indicate the occurrence of a word in the documents of a specific organization, and 0 otherwise); (2) the use of a lexicon (WordNet) to convert words into hypernyms or supersets of words to capture the concepts expressed within the organization’s webpages; (3) development of decision-rules for the retention of hypernyms that are within the topic of interest to us, by the use of a-priori specification of hypernyms of interest, e.g., “computation”, “platform”, “hacking” and so on; a total of 40 hypernyms were retained for further analysis; (4) determining the weight given within an organization’s documents to a concept by calculating the percent occurrences of that hypernym out of the total of all hypernyms—this is denoted by the variable name hypernym_weight; (5) clustering organizations based on the hypernym occurrence percent, using a k-means clustering method, where an Iteratively Reweighted Least Squares minimizes the root mean square difference between the data and the corresponding cluster means.

Various metrics were used to determine the final choice of number of clusters and spatial disjointness of the clusters, including the overall $R^2$ of 0.49 and Sarle’s Cubic Clustering Criterion (CCC) of 1.2, lending further evidence that the clusters are spatially separated enough to provide meaningful groupings; and (6) labeling clusters according to the relative values of the hypernym_weight within the cluster.

### 3. Analysis of Urban Digital Infomediaries

In our approach, there are three aspects to understanding the emerging urban data sector: identifying the types of entities that are active in this space; understanding the scope of what the organizations do; and, assessing the extent to which these boundaries are blurring over time towards the goal of inferring trends towards convergence. We identify four major groups of urban digital infomediaries consisting of 9 specific organizational types, based on their stated mission and objectives, and the products and services delivered. Table 1 shows these four groups, with description of specific types of organizations within each group, and the sectors (public, private, non-profit, informal) to which they belong. The table also displays the percentage of the total sample which a specific type of organization comprises of.
Table 1: Urban digital infomediaries and dominant sector

<table>
<thead>
<tr>
<th>Organization Type</th>
<th>Description</th>
<th>Dominant Sector</th>
<th>% of Total Sample (N=139)</th>
</tr>
</thead>
<tbody>
<tr>
<td>General-purpose ICT Infomediaries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCC Smart City Companies (includes units of comprehensive ICT businesses)</td>
<td>Companies or business units focused on improved performance and efficiency of city ICT systems</td>
<td>Private</td>
<td>18</td>
</tr>
<tr>
<td>MSICTC Multiple-service ICT Companies</td>
<td>Organizations providing multiple hardware, software, and communications services targeted to location-based information, information-sharing, collaborative tools and related products</td>
<td>Private</td>
<td>6</td>
</tr>
<tr>
<td>Urban Information Service Provider Infomediaries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CIS City Information Services</td>
<td>Organizations providing directory services or other information for residents to connect to social, entertainment and other aspects of cities</td>
<td>Private, Informal</td>
<td>4</td>
</tr>
<tr>
<td>LBS Location-Based Services</td>
<td>Organizations providing generic location-focused services including navigation, retail, health and wellbeing based on location as well as social interaction and urban engagement opportunities based on location-based social networks</td>
<td>Private, Informal</td>
<td>21</td>
</tr>
<tr>
<td>Urban Open and Civic Data Infomediaries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ODO Open Data Organizations</td>
<td>Organizations publishing open data for further analytics and use</td>
<td>Public, Non-Profit, Informal</td>
<td>30</td>
</tr>
<tr>
<td>CHO Civic Hacking Organizations</td>
<td>Organizations analyzing and distributing civic statistics, maps and other information of interest for civic and public discourse</td>
<td>Non-profit, private, informal</td>
<td>3</td>
</tr>
<tr>
<td>CBISO Community-Based Information Service Organizations</td>
<td>Organizations and individuals connecting information services to specific cities, communities and neighborhoods through analytics, content creation, visualization, mapping and other methods</td>
<td>Public, Non-Profit, Informal</td>
<td>11</td>
</tr>
<tr>
<td>Independent and Open Source Applications, Software and Content Developer Infomediaries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IAD Independent App Developers</td>
<td>Individuals primarily focused on developing software and apps to link citizens to information</td>
<td>Informal</td>
<td>3</td>
</tr>
<tr>
<td>OSD Open Source Developers</td>
<td>Organizations and entities creating open source software, social coding accounts, developer networks and other open source ways to allow access to Big Data and Open Data</td>
<td>Private, Informal</td>
<td>4</td>
</tr>
</tbody>
</table>
To a certain extent, the distinctions we have drawn among organizations is already product or service-based. Nevertheless, it is useful to examine the range of products and services with which urban data organizations are involved and the extent to which organizations rank high or low in being a producer versus a user of specific products and services. Four types of product-service mix can be identified for the organizations examined: (1) data generation, communication and management technologies such as sensor systems, wired and wireless communication systems, information processing and database management systems, positioning systems, web services, and associated hardware and software to manage data, (2) software solutions including data platforms and tools to enable further use of urban data; (3) analytics and knowledge-discovery services including data mining, urban and regional modelling, mapping and visualization and human interpretation of the results, towards understanding and exploring cities and communities, and (4) end-user services including collaborative community decision support tools, user apps for numerous functions and various web-based information services. These are described in Table 2 with our qualitative ranking of the infomediaries on the four product-service mix considered, with “H” indicating our assessment of an organizational type ranking “High”, “M” for “Medium” and “L” for “Low”.

Table 2: Organizational Products and Services

<table>
<thead>
<tr>
<th>Infomediary Type</th>
<th>Data Management and Communication Tools</th>
<th>Platforms &amp; Tools to Connect Data to Communities</th>
<th>In-house Analytics and Knowledge-Discovery</th>
<th>End-User Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCC</td>
<td>H</td>
<td>H</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>MSICTC</td>
<td>H</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>CIS</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>H</td>
</tr>
<tr>
<td>LBS</td>
<td>H</td>
<td>L</td>
<td>M</td>
<td>H</td>
</tr>
<tr>
<td>ODO</td>
<td>H</td>
<td>H</td>
<td>M</td>
<td>L</td>
</tr>
<tr>
<td>CHO</td>
<td>L</td>
<td>M</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>CBISO</td>
<td>L</td>
<td>L</td>
<td>H</td>
<td>M</td>
</tr>
<tr>
<td>IAD</td>
<td>M</td>
<td>L</td>
<td>M</td>
<td>H</td>
</tr>
<tr>
<td>OSD</td>
<td>H</td>
<td>M</td>
<td>L</td>
<td>L</td>
</tr>
</tbody>
</table>

Based on the information above, the urban ICT infomediaries are as follows:

**General Purpose ICT Infomediaries**: These organizations are most likely to provide services towards building and managing the intelligent infrastructure in networked cities, and is composed almost entirely of private firms. The primary business models for this group are Business-to-Government (B2G) and Business-to-Business (B2B), although the value of their foundational information infrastructure ultimately benefits myriad of end users. The group is composed of two types of business organizations which are distinguished by the extent of urban emphasis in their overall product–service mix (for example, the percentage of total business geared to urban ICT products and services).

1) **Smart City Companies (or Units)**: These entities are focused on improved performance and efficiency of city ICT systems. While there are many definitions of smart cities, the overall vision is one of having ICT-focused solutions be an integral part of urban development, driving economic
competitiveness, environmental sustainability, and general livability. As noted by Thakuriah and Geers (2013), the term “smart city” is championed by commercial entities and the expectation is that the networking and integration of multiple urban sectors will enable cross-agency efficiencies for a range of services (such as traffic management, utilities, law enforcement, garbage disposal, emergency services, aged care, etc). In some cases, the entire SCC business is focused on city-centric applications of intelligent infrastructure, data management and analytics, and in other cases, the smart city business is handled by specific units within comprehensive ICT businesses which are involved in many other ICT sectors such as health, finance, energy and so on. This group also includes consulting firms which have historically offered services in Intelligent Transportation Systems, smart energy and water management and other sectors. As shown in Table 2, SCCs are high on data management and communications tools and in developing platforms and tools for further processing of data. They are likely to be involved in analytics and knowledge discovery processes for the purpose of business solutions, but are likely to be involved in end-user services to a lesser extent.

2) Multiple-Service ICT Companies: These are business organizations providing foundational, general-purpose hardware, software, and communications services targeted to location-based information, information-sharing, search engines, map databases, web services, sensing technologies, collaborative tools, social media, Web 2.0 and related products. They provide general ICT services in the telecommunications and information technology sector, without being focused on urban applications such as smart cities, or are focused on them only in incidental ways, without making the urban focus a core aspect of business; yet, the information infrastructure they provide are vital to urban data initiatives. For example, they are owners of cell phone data which has a wide variety of urban mobility analysis applications. As in the case of SCC, MSICTC can range from small private firms to large multinationals.

Urban Information Service Infomediaries: The second group consists of organizations which deliver end-user ICT services to urban residents to explore cities and communities, and to connect citizens to social, entertainment, economic and commercial opportunities. The services provided are connected to specific business models such as mobile commerce and location-based advertising or banner ads. While the majority of the organizations we examined are established private businesses, some of these services are also being offered by informal, independent developers. Two specific digital infomediaries can be differentiated by the types of information they provide, the degree to which location and real-time information streams are explicit and front-and-center in their products, and the extent to which social networking processes are utilized.

3) City Information Services (CIS): CIS organizations provide directory services, question and answer databases or recommender systems for residents to engage in social, commercial, entertainment and other aspects of cities. As shown in Table 2, CIS are generally focused on end-user services. These organizations often utilize crowdsourced information by means of user reviews and ratings of businesses, entertainment services, restaurants and other retail and commercial entities, thereby creating user communities who may form a social network. CIS infomediaries tend to be private companies or informal organizations and independent developers. They are distinguished from Location-Based Services by not requiring explicit positioning and navigation capability, which call for additional (positioning and sensor) technologies.

4) Location-Based Services (LBS): The LBS industry is either private and to a limited degree, informal, and has been studied extensively. LBS are information services that capitalize on the knowledge of, and are relevant to, the mobile user’s current or projected location. Examples of LBS include resource discovery services in response to spatial queries such as “Where am I?” or “What’s around me?” Other examples are directory assistance and service location (for example, find the nearest gas station with cheap gas), Points of Interest locations (for example, find the social services building), routing and navigation and many others. The LBS industry has been noted to be highly heterogeneous, with many different types of players. These include mobile operators, content providers, content aggregators, wireless application service providers. One specific group of LBS
are Location-Based Social Networks (LBSN) which explicitly connect location to social networks not only by adding location information to an existing social network so that people in the social structure can share location-based information, but also new, perhaps ad-hoc social networks made up of individuals connected by locational proximity (Zheng, 2011). Due to their utilization of location-aware technologies, real-time and heterogeneous information sources, LBS companies are likely to be high on data management aspects, as well as end-user solutions.

**Urban Open and Civic Data Infomediaries:** This group consists primarily of government, non-profit and informal entities, but with a sprinkling of private organizations (in this highly dynamic environment, it is also possible that the informal entities are looking for business models and private equity funds). The major goal of this group, which consists of three types of infomediaries, is to make government data available for further use or to work with administrative and other data to enable public discourse regarding government transparency, community involvement and civic engagement.

5) **Open Data Organizations (ODO):** These are primarily ICT-focused units within government agencies publishing Open Data for further analytics and use. Many cities now have data portals where government agencies upload digital information that is license-free and in non-proprietary formats. Data that are being released in portals are on transportation, public, health, crime, public works and public services, education, and economic development investments and business affairs. Recent trends in government-supported APIs and mashups have enabled the ability to tap into such data in a growing number of ways. ICT entrepreneurs in government agencies and civic leaders pushing the agenda of open standards and open source software has been catalysts in Open Data initiatives. Table 2 shows that ODSs are also likely to be involved in data management and communications aspect, and due to their objective of making data accessible for end-uses, also on platforms and tools for further use.

6) **Civic Hacking Organizations (CHO):** CHOs are involved in data-centric activism relating to civic and community issues of interest to them, through analytics, visualization and knowledge-discovery of administrative data. This information is shared as civic statistics, maps and other media with government and community leaders, other decision-makers and the general public, thereby generating informal urban analytics on civic or governance matters of value to specific user communities. Within this sphere of work, informal analysis relating to government transparency, governmental funding priorities, and related topics have gained much attention. As noted previously, in addition to being organized into CHOs, civic hackers may act independently, without being a part of a formal organization. CHOs may also be involved to a lesser in the creation of collaborative, interactive solutions to present data for additional processing by others.

7) **Community-Based Information Service Organizations (CBISCO):** These are typically established community organizations providing a range of services in addition to connecting data and information services to citizens of specific communities and neighbourhood. The information that is disseminated include the output of analytics, content creation, visualization, mapping and other methods. CBISOs are, like CHOs, highly involved in analytics about their area, but the analytics work may not be done by the organization itself; their value in the urban data chain is to filter and package appropriate information for use by their stakeholders. They are also likely to host community decision-support and collaborative tools for participatory neighbourhood problem-solving.

**Independent and Open Source Applications, Data and Content Developer Infomediaries:** This group of ICT-focused infomediaries develop cross-cutting urban applications and user apps but can be distinguished from the previous groups in that their work takes place in primarily independent and informal ways, in contrast to established organizations.

8) **Independent App Developers (IAD):** This group is composed of entrepreneurial ICT developers, who are primarily focused on developing apps to link citizens to information primarily in the Business-to-Customer (B2C) sector. Ferraro et al (2011) have identified several approaches to
building user communities and revenue models for B2C commerce, which are particularly relevant to IADs, including “freebie”, advertising-supported, and “premium” versions. IADs are involved in end-user solutions, and in the process of doing so, they may be a part of data management and analytics solutions.

9) Open Source Developers (OSD): Of great value to the open data community are open source software and social coding accounts, which makes data, coding and analytics tools freely available. There has been a general shift in the urban data communities to open standards, software and resources. For example, major Big Data management tools such as Hadoop and Pig are open source. Urban-focused OSD develop or modify open source software for the creation of civic statistics, contribute lines of code to social coding accounts, and develop open-source developer API. OSDs also contribute to urban data through developer networks and other networked approaches to allow access to urban data. OSDs are likely to be a part of data management and communications activities, as well as activities relating to the development of solutions for further processing and analytics of data by others.

3. Activity Clusters: Approaches and Processes for Convergence

One of the questions raised earlier is the potential blurring of activities among different types of infomediaries with different mission/objectives. The text mining exercise described in Section 2 was used to extract information on the types of activities undertaken by the 139 organizations.

This led to the identification of 7 Activity Clusters given in Table 3 which describe processes and approaches undertaken to operationalize organizational objectives: (1) data, computation and tool-building; (2) accessible, advisory, citizen-oriented; (3) economically efficient and resilient urban communities; (4) ICT-focused urban systems management; (5) smart and sustainable communities; (6) community information and location services; and (7) accountability, advocacy and data activism.
### Table 3: Activity Clusters of Urban Digital Infomediaries

<table>
<thead>
<tr>
<th>Activity Cluster No.</th>
<th>Activity Focus Cluster Label</th>
<th>Description of activities</th>
<th>Exemplar hypernames with high ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Data, computation and tool-building</td>
<td>Involved in intelligent infrastructure and computationally-intensive application development</td>
<td>Computational and data management constructs such as “algorithm”, “code”, “intelligent”, “computation”, “platform”</td>
</tr>
<tr>
<td>2</td>
<td>Accessible, advisory, citizen-oriented</td>
<td>Work include a focus on citizen-oriented accessibility and equity concerns, possibly on the social justice aspects of the use of ICT</td>
<td>“accessibility”, “advisory”, “citizen”, “equity” and related terms</td>
</tr>
<tr>
<td>3</td>
<td>Economically efficient and resilient urban communities</td>
<td>Focus on ICTs from the perspective of economic development, resiliency and sustainability</td>
<td>“urban”, “communities”, “neighborhood”, “economic efficiency”, “effectiveness”, “collaborative”</td>
</tr>
<tr>
<td>4</td>
<td>ICT-focused urban systems management</td>
<td>Focus on urban systems management, but with some urban engagement, e.g., technology for resource discovery, community monitoring, and sharing</td>
<td>“system”, “efficiency”, “management”, “urban”, “intelligent”, “engagement”</td>
</tr>
<tr>
<td>5</td>
<td>Smart and sustainable communities</td>
<td>Focus not unlike Clusters 2 and 3, but with an additional focus on smart and sustainable urban communities</td>
<td>“accessibility”, “advisory”, “citizen”, “communities”, “economic efficiency”, “smart”, “effectiveness”, “collaborative”</td>
</tr>
<tr>
<td>6</td>
<td>Community information and location services</td>
<td>Involved in community-based information and ICT support</td>
<td>“community”, “apps”, “location”, “collaborative”, “participation” and “platform”</td>
</tr>
<tr>
<td>7</td>
<td>Accountability, advocacy and data activism</td>
<td>Focus on data-centric activities for civic activism and advocacy</td>
<td>“accountability”, “transparency”, “activism”, “participatory” and “collaborative”</td>
</tr>
</tbody>
</table>

The spread of the 9 different organizational types across these 7 Activity Clusters is given in Figure 1. Given the overall focus on technology, it is not surprising that all 9 organizational types have at least some organizations that are dominant in Activity Cluster 1, “Data, Computational and Tool Savviness”. However, for close to 35% of ODO organizations, this is the dominant activity, as determined by hypernyms, followed by LBS and MSICTs, indicating that such organizational types with their interest in data sharing and publication are pushing the envelope in technology development around urban data.

**Figure 1: Distribution of organization type across activity clusters**

![Figure 1: Distribution of organization type across activity clusters](image-url)
LBS and SCCs dominate in the “ICT-focused urban systems management cluster”. In the case of this activity cluster, as in the case of first cluster, it is somewhat surprising that Community-Based Information Service Organizations dominate quite strongly in terms of overall focus, indicating that very different types of organizations are carrying out functionally similar activities. The dominance of SCC, ODO and CBISO organizations within the “Accessible, advisory and citizen-oriented” activity cluster similarly suggests increased convergence in the use of data and technology to serve cities.

Similar trends regarding the dominance of very different types of organizations within other activity clusters seem to indicate increasing convergence of focus. Studies of convergence go back to the analysis of the machine tool industry by Rosenberg (1963) and recently, the ICT sector has received considerable interest – convergence not only in technologies from the supply side used but also in products from the demand side (Stieglitz, 2003). It is likely that as urban digital infomediaries in historically different industries increasingly use urban data to produce similar products and services, convergence may be stimulated, leading to potentially useful networks, alliance strategies and partnerships.

4. Skills and Interests of the Workforce Involved and the Role of Networks

Within these organizations are a wide spectrum of professionals with interests in urban data and related analysis and operations, and their background and experiences partly shape the activities of the infomediaries. One group of professionals is “data scientists”. One of many definitions of data scientists is that they are individuals “with the training and curiosity to make discoveries in the world of Big Data” (Patil and Hammerbacker, n.d.). Data scientists are likely to be drawn into urban analytics due to the data-rich characteristic of cities, and the potential for scientific discoveries and commercial innovations.

A second group which is involved in urban data are ICT developers. Developers perform myriad functions within this environment, including the integration of ICT systems of different urban sectors, development of city or community dashboards or “apps”. This subgroup may also include ICT-oriented individuals social entrepreneurs (civic hackers) who produce curated information including mashups that convey relevant events and apps for others to use (“apps for democracy”) and ordinary citizens who access the data to understand more about transportation and other conditions in their communities.

A third group consists of analysts trained in urban and regional planning or related disciplines such as geography, economics, or civil engineering. Some within this subgroup are professionals with an interest in urban models and simulations for quantitative analysis of urban areas, but in contrast to the models of data scientists, which tend to be “data-driven”, analytical tools of urban and regional planning professionals are driven by a long history of various aspects of urban theory (for example, regarding urban economics, mobility and so on) and in the quantitative analysis of various aspects of cities, such as transportation, regional economic analysis, public health, environmental planning and so on, and in the theoretical and methodological aspects of their subjects.

A fourth group consists of managers who lead data-centric projects in cities or communities. Examples include private sector managers responsible for delivering a Big Data sensing project to a government agency or public managers responsible for bringing Open Data Portals online. Other examples include community planners who are establishing neighbourhood dashboards with community information and civic statistics, or transit managers who are responsible for delivering a real-time transit arrival system. Urban data project managers may be required to take on multiple leadership roles in order to see a project from start to finish, including project championing (Lam, 2005), having the social skills to communicate sometimes highly technical information to non-technical audience, as well as technical skills, team building and project management skills (Gil-Garca, et al, 2005).

General Purpose ICT Infomediaries are most likely to have developers, data scientists and ICT managers and administrators in their mix of employees. The Urban Information Service Provider Infomediaries group too is likely to be staffed with developers and data scientists. They are also likely to utilize the work of citizens who generate information via sensing and crowdsourcing systems, which are then shared with other
service users. Urban Open and Civic Data Infomediaries are likely to employ personnel formally trained in disciplines in the urban and regional planning domains, whether as methodologists or generalists interested in civic and community issues. Within this group, informal civic hackers with no formal ties to any of the organizations discussed here may be actively involved. This group is also likely to utilize the work of ordinary citizens in the monitoring and reporting of events within their communities.

Independent App Developers or Open Source Developers are very likely to be developers or data scientists. It is also possible that within the IAD group, citizens with no formal training in informatics or the urban disciplines, but who have self-taught the use of social software and open source tools, are playing a part. Data-centric managers, as described above, are likely to be employed in all infomediary groups considered, since broad-based ICT technical, managerial and entrepreneurial skills are needed in virtually all urban data sectors.

5. Role of Networks

As noted previously, many of the activities driving urban data are not occurring within organizational boundaries, but through formal and informal networks involving the actors discussed above. We discuss this aspect here very briefly, with the note that this topic is a significant research area in its own right.

Informal networks of informed citizens, civic technologists and civic hackers have become an important aspect of urban data and it is possible that they are attracting individuals from a wide variety of ICT organizations, although to the best of our knowledge, we have not seen the results of any study on this particular topic. In the area of civic hacking, ongoing networks are emerging due to the use of social networking (both online and face-to-face meetings via Meetup groups) to exchange knowledge among civic hackers who are more tech-savvy developers and the less technically savvy, and to discuss developments in software, data and policy and civic issues, is becoming increasingly important. “Hackathons” that are being sponsored by government agencies and non-profits, as well as design and crowdsourced competitions for citizen apps, are giving increasing identify, visibility and legitimacy to these activities. A slightly different type of networks are those spearheaded by primarily established ICT companies, and focused to a greater degree on urban data and urban management applications, in contrast to civic and government transparency issues.

As indicated earlier, members of the organizations we discussed tend to populate ongoing networks relating to data, communication and other standards, and other technical aspects related to urban data. In contrast to ongoing networks, there are also project-based networks in the urban data sector. These are formed primarily around technology-focused city projects (smart city projects, field operational tests of intelligent transportation, smart energy systems and so on) primarily by members of government agencies who are sponsoring the project, businesses that are providing the service, and affiliate members consisting of other planning and administrative agencies, non-profits, higher education and research institutions. These types of networks are often formed as result of project requirements for partnerships or public participation, and their work typically end when the project is complete, although they may continue to come together well after that to follow up on evaluation results and to develop strategies regarding lessons learned.

6. Urban Data Infomediaries and Functional Dimensions

In the previous discussion, several different measures were considered in understanding urban data infomediaries: ie, their mission, products and services, major activity clusters, and type of workforce involved. Based on the information contained in the measures, urban data infomediaries can be considered as having five dimensions of functional interest: techno-managerial, scientific, business and commercial, urban engagement, and openness and transparency. The alignment between the organizational type and the dimension of functional interest is given in Table 3.
Table 3: Functional Interest Dimensions of Urban Data Infomediaries

<table>
<thead>
<tr>
<th>Infomediary Type</th>
<th>Functional Interest Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smart City Companies</td>
<td>Techno-Managerial; Business and Commercial; Scientific</td>
</tr>
<tr>
<td>Multiple-service ICT Companies</td>
<td>Business and Commercial; Scientific</td>
</tr>
<tr>
<td>City Information Services</td>
<td>Business and Commercial; Urban Engagement</td>
</tr>
<tr>
<td>Location-Based Services</td>
<td>Business and Commercial; Techno-Managerial; Scientific</td>
</tr>
<tr>
<td>Open Data Organizations</td>
<td>Openness and Transparency; Urban Engagement; Techno-</td>
</tr>
<tr>
<td></td>
<td>Managerial; Scientific</td>
</tr>
<tr>
<td>Civic Hacking Organizations</td>
<td>Openness and Transparency; Urban Engagement; Scientific</td>
</tr>
<tr>
<td>Community-Based Information Service Organizations</td>
<td>Urban Engagement; Techno-Managerial; Urban Engagement</td>
</tr>
<tr>
<td>Independent App Developers</td>
<td>Business and Commercial; Urban Engagement; Scientific</td>
</tr>
<tr>
<td>Open Source Developers</td>
<td>Openness and Transparency; Scientific</td>
</tr>
</tbody>
</table>

One dimension is *techno-managerial* and focuses on how urban data can foster effective management of cities through interconnectivity among different urban sectors, collaboration with citizens and communities, and dynamic resource management. A second dimension of interest is *scientific* focusing on quantified urbanism with the view that urban data can help learn about cities in new ways thereby creating new scientific understanding and knowledge discovery; however, scientific does not mean a purely research-level endeavor, as scientific discoveries can be put to operational and policy use. This dimension also views data on city dynamics as offering interesting opportunities and test-bed to address communications, information processing, data management, computational, and analytics challenges posed by urban Big Data. Another aspect of the scientific interest is to make advances in open source and social software, technologies for privacy preservation and information security, and other challenges associated with urban data.

A third dimension is *business and commercial* where previous modes of e-commerce are being augmented with location-based social networks for mobile commerce, user-generated content for reviews and recommender systems, crowdsourcing of input for idea generation and other business product development, and other commercial purposes in cities, ultimately leading to participant-generated information on the social, recreational and entertainment aspects of cities.

A fourth dimension of interest is *urban engagement and community well-being*, with a focus on civic participation and citizen involvement. One aspect of this strand is community-based information and monitoring, which may involve technologies similar to techno-managerial strand.

A fifth dimension is *openness and transparency* of government information towards more interactive, participatory urban governance and bearing close similarities to other “open” movements including open access, open source, open knowledge and others. One aspect of this strand of interest is likely to overlap with the second aspect of the scientific dimension, ie, on computational and data management aspects.

7. Conclusions

Our objective was to make a qualitative assessment of public, private, non-profit and informal infomediaries who are pushing the agenda of the urban data sector. Using a mixed-methods approach, we identified four major groups of organizations: general-purpose ICT infomediaries, urban information service provider infomediaries, urban open and civic data infomediaries, and independent and open source developer infomediaries.

A total of nine organizational types were highlighted within these four groups. Organizations are found to have the following seven areas of focus regarding their activities and process: data, computation and tool
development; accessible, advisory, citizen-oriented; economically efficient and resilient urban communities; urban systems management; smart and sustainable communities; community information and location services; and accountability, advocacy and data activism focus. A variety of professionals are involved within these organizations in urban data including urban and regional planning professionals, data scientists, developers, and ICT-trained or urban-trained project managers. The work of the organizations also involves others such as civic hackers and citizen scientists, in a mixed of paid work and volunteer efforts.

The urban data sector is highly dynamic and involves a significant informal entrepreneurial sector which is entering the domain given the opportunities and the overall challenges involved. However, for these opportunities to be realized, a broad-based strategy is needed that reflects research and policy deliberations regarding the social, policy, behavioral and organizational implications of the data management and dissemination processes. This entrepreneurial sector is itself highly diverse and includes developers passionate about computational and technological challenges, and data scientists who are interested in analyzing complex data, as well as in open source software.

Urban data are also increasingly being seen by professionals in the urban planning and public management domains as being important towards urban management. It also includes the work of civic hackers and civic technologists who value openness and transparency and open source technologies, and are interested in data and analytics to address civic and urban challenges. While some entrepreneurs work in the private firms, civic and public agencies, or in non-profits, others freelance in ICT development work.

Using all the different measures, the dimensions of interest addressed by urban data infomediaries are five-fold: techno-managerial, scientific, business and commercial, urban engagement, and openness and transparency. Using these dimensions, it may be possible to predict aspects of city operations and management where innovations may result from the work of urban data infomediaries.

Informal networks are playing an important role in creating and sharing knowledge regarding technical skills and urban and governance issues. Social networks that have formed among developers, data scientists and others involved in civic hacking are important in this regard, but requires greater involvement of professionals from the urban research community. The participation of the latter group is particularly important, since what is sometimes presented as novel digital modes of urban planning, are effectively practices and strategies that are well-established in the urban domain. In much the same way, much can be learned about emerging technology and analytics solutions by the urban community. This indicates that there should overall be greater cross-fertilization of knowledge among the various professional domains involved.

One issue we were interested in is the idea of convergence, which is a general trend in the ICT sector, as noted by many authors. We found evidence of technical convergence because many different types of organizations use similar technologies and offer similar urban data products and services. We also found through our clustering analysis that several different types of organizations across the four infomediary groups are focused on similar ICT-focused activities relating to urban management, accountability and data activism.

However, less evident is ideological convergence. Ideological tensions occur with differences in viewpoints regarding the way things should be, or should be done. There are several examples in the urban data landscape. A far from complete list is on ideological tensions regarding what open government should mean: transparent government, innovative or collaborative government, or in the case of the example given earlier, should there be a distinction between physical urban planning and digital urban planning? These questions are important to consider because they have implications for policy and stakeholder generation, as well as for practice and bodies of knowledge in these areas.

The study has several limitations. The emerging nature of the sector necessitated an exploratory study. Aside from the informal nature of the sample of organizations examined, another limitation of our sample consist of only those entities that have a formal presence (websites) in the Internet. Informal digital infomediaries who do not have websites but are active through blogs, social networking sites such as Facebook, or have a social web presence via social coding services such as Github, SourceForge and so on,
or who have no presence in the Internet at all, are not included in this study. This potentially excludes a significant share of informal digital infomediaries including civic hackers, many of whom are one-person entities contributing without an established organizational presence. However, we were able to identify some organizations which are undertaking civic hacking activities and these are included in the sample. At the time of writing this paper, we are administering a survey instrument to gather data on such the independent data activists and civic hackers. Another limitation of the sample is that it excludes higher education and research institutions, some of which are focusing heavily on Big Data and Open Data research.

References

The Influence of Distance Decay Coefficient on Modularity-based Community Detection in Space-embedded Networks

Yaoli Wang 1*, Song Gao 2

1 Dept. of Geography, University of Georgia, Athens.
2 Dept. of Geography, University of California, Santa Barbara.

*Corresponding author: wyl0810@uga.edu

Abstract

In urban studies, it is important to know which places (e.g., functional regions) are more closely interacted than others regarding human movements, freight traffic and information communication. Modularity-based community detection is a method to find network clusters (i.e., communities within which interactions are stronger than across) by maximizing modularity, a measure of the gap between the observed and the expected interactions. We involve distance decay model to approximate the expected interaction in space-embedded network. It is supposed that the distance decay coefficient has a crucial influence on the detection result: as in gravity model, if the decay coefficient for observed interaction is smaller than that for expected interaction, namely, if actual interaction has more long-distance trips than expected, we will observe significantly more exclaves (i.e., places not geographically adjacent but belonging to the same community). With the geographic and demographic information of cell phone towers in a city of China, we apply different distance decay coefficients and generate totally 6 groups of interactions as the “observed”
interactions. By adding the same series of variations (including positive and negative ones) to each of the observation group’s coefficient, we get the coefficients for the “expected” interactions. In this way, we simulate scenarios when the observed interaction decays slower or faster than the expected. The results of hypothesis test support our original assumptions on the emergence of exclaves.

Introduction: Community Detection and Exclaves

With the widespread use of location-awareness devices and sensors such as mobile phones and GPS, a great opportunity emerges to study complex human-environment spatial interactions based on such urban big data. In urban studies, it is important to know which places (e.g., functional regions) are more closely interacted than others regarding human movements, freight traffic and information communication. This may yield insights for urban planners on land use patterns and for transportation engineers on reasons of congestions. However, traditional spatial clustering approaches are not sufficient to explore network structure of spatial interactions.

Network analysis was initially used in biology and sociology, but contemporarily there emerges a trend of network analysis in space-embedded context [1]. As diverse methods and measures of network analysis can contribute to different properties of a network, community detection is helpful in figuring out closely related clusters and facilitating deeper understanding into the interaction between nodes of network. Community detection is the method and process to find out clusters (i.e., communities) in networks such that interactions within the same community are stronger than across communities. When being introduced into space, community detection helps to find groups of places with strong spatial interaction (such as human movement flow,
commodity flow, currency flow, and information flow in mobile phone communications). Spatial community detection shares the similar idea of grouping but differs from spatial clustering which mainly considers characteristics of attributes and spatial distribution in spatial data.

A major community detection method is modularity-based community detection. Modularity was firstly proposed by Newman and Girvan[2], defined as the difference of observed interaction (OI) strength in the studied network beyond the expected interaction (EI) strength in a random network (eq. [1]) to testify the significance of detected community.

$$Q = \frac{1}{2m} \sum_{C \in p} \sum_{i,j \in C} |A_{ij} - P_{ij}|$$  \hspace{1cm} [1]

where $A_{ij}$ is the observed interaction between two places (or regions) $i$ and $j$, $P_{ij}$ is the EI, and $m$ is the sum of link weights all over the network.

Though community detection is no more a novel topic, community detection in space-embedded network is not yet sufficient [3–6]. This is because the spatial context affects the structure and the properties of the spatial networks, including (1) spatial constraints on the connectivity of space-embedded nodes; (2) physical networks like roads and railways, which are affected by spatial topology; and (3) restrictions on long-distance links due to temporal or economic costs. It is, therefore, crucial to know what the role is of the spatial context in community detection. In this study, we concentrate on the influence of the geographical distance.

Originally, some work (e.g., [6]) simply introduced modularity directly into spatial network without modifying the formula of modularity, and got the result that,
even without any constraint on geographical adjacency, spatially clustered communities could be well detected, corresponding to the concrete administration boundaries. However, such finding is not a big surprise because of the common geographic phenomena, distance decay. According to Tobler’s First Law [7], “Everything is related to everything else, but near things are more related than distant things”, so it is normal to find spatial clusters emerging spontaneously.

To take into account the nature of space, researchers (e.g., [3]) have involved the distance decay model, usually in the form of gravity model, into modularity as the approximation to the EI (i.e., $P_{ij}$ in eq.[1]) in order to get rid of the constraint of distance. The gravity model has the general formula as equation [2], where $k$ is a constant coefficient, $N_i$ and $N_j$ represent the mass of two places (usually representing population, GDP, or an integrated index), and $f(d_{ij})$ is the decay function negatively related to the distance between $i$ and $j$. As a result, there appear more exclaves. Exclaves are the units (i.e., the basic study area representing a place) that belong to the same community but not spatially neighbored. Note that if some units of a certain community can be connected via some other units of that community, these places are called “islands”, instead of “exclaves”. If there are more than one island belonging to the same community, they generate an exclave. In this work, we quantify the amount of exclaves of a detection result by the average number of islands in each community. We will show an example later on. The exclaves denote the units between which the interactions are stronger than the average expected by the distance decay model on that distance. Hence they indicate long-distance interactions.

$$P_{ij} = k N_i N_j \left(f \right)$$

[2]

However, when deciding the specific formula of the distance decay model, we
note that the distance decay coefficient plays an important role in determining the
pattern of exclaves. When adopting the traditional power law function (eq. [3], where
$k$ has the same meaning as above, and $b$ is the distance decay coefficient) to
approximate the EI, we get the pattern shown in Figure 1-a. But when utilizing the
deterrence function (eq. [4], where $N_i$ and $N_j$ stand for the same as above and $A_{ij}$ is the
observed interaction between two nodes) proposed by Expert et al. [3], the pattern is
different as shown in Figure 1-b. The deterrence function is drawn directly from
empirical data without any prior knowledge on the particular format of distance decay,
and can be converted approximately to a power law function (SI text A).

$$f(d_{ij}) = \frac{k}{d_{ij}^b} \quad [3]$$

As shown above, if the EIs in both cases can be quantified by power law function,
the reason for the apparent difference is likely to be the different distance decay
coefficient. We observe a slight difference in the value of decay coefficient $b$ between
fitting distance decay function with power law ($b = 1.65$) and transforming the
deterrence function of eq. [3] into power law ($b = 1.59$). This slight dissimilarity
surprisingly yields a big difference in the detected result: the former case has more
exclaves (Figure 1-a), whereas the latter looks random in space (Figure 1-b). We are
thus inspired to discuss the influence of distance decay coefficient on the detection
results.
Figure 1. Detected communities after involving distance into modularity, without compulsory requirement on spatial adjacency. The numbers of communities in a) and b) are the same. The colors are randomly assigned, and the same color represents the same community. Cells in the same color (community) but not spatially connected by other cells of that color form exclaves. For instance, visually apparent, there are four big exclaves in a), in brown, in light brown, in yellow, and in dark purple.

Exclaves implicitly signifies the uneven spatial distribution of urban resources.
By zooming to exclaves, urban planners might be able to find the reasons for the strong attractions between certain land-use types, which may be potential causes for unwanted congestion. We focus on discussing the influence of distance decay coefficient in space-embedded community detection to deepen our understanding into the nature of the city in terms of human movement behavior and implicit land use types.

**Distance Decay Coefficient in Community Detection**

The main idea of this work is to discuss the influence of decay coefficient $b$ on the detected exclaves by an abstract control experiment. The experiment simulates a series of “observed interactions” as the interaction flows with a series of pre-defined decay coefficients. By adding a series of variations to each of the coefficients for observed interactions, we get the coefficients for EIs. Based on that, we perform community detection in our simulated scenarios with background information of mobile phone tower service area (“cell”) in a city of China. More detailed information about the geographical background and data processing can be found in the previous work [8]. A cell is a unit (see above Figure 1) in this study. We will demonstrate the change of detected results under different scenarios, i.e., relationships between decay coefficients of the expected and of the observed interactions. The study area is a city\(^1\) in the northeast part of China. A cell (namely, the area covered by each mobile tower) is generated as the Voronoi diagram around that mobile phone tower. Each cell has population information and pairwise phone call volume to all other cells in a certain period of time. However, since we just need to simulate interaction flow between each pair of cells, the call volume is not really utilized; all the call volumes in this study are

---

\(^1\) As requested by the data provider, the name of this city is not revealed.
Instead of adopting the modularity as equation [1], we slightly revise the formula of modularity:

\[
Q = \sum_{C'P_{i},j\in C} \left| \frac{A_{ij}}{P_{ij}} \right| = \sum_{C'P_{i},j\in C} \left| \frac{A_{ij}}{N_{i}^{m}N_{j}^{m}d_{ij}} \right|^\varepsilon \tag{5}
\]

where all the notations have the same meaning as aforementioned. There are two things worth attention. First, we incorporate gravity model into modularity so as to consider the essential effect of distance decay on the EI strength. The spatial adjacency of community detection by Great Britain’s landline [6] is primarily due to distance decay; otherwise more exclaves will appear as in [3]. Since we want to examine the long-distance travel beyond the impedance of space cost, it is rational to involve gravity model. Second, as argued in another paper [8], we adopted division form so as to detect the percentage change to allow for the bias of cell size. Larger cell will intrinsically have higher flow, resulting in bigger difference in absolute value. This is different from Expert et al.’s work [3] who adopted subtraction, despite we all take gravity model into account.

We can transform equation [5] further to make the influence of decay coefficient explicit. Many observations have demonstrated the existence of distance decay in real geographical phenomena [8-11], such as cell phone calls, urban movements, to name a few. Therefore, we argue that the observed interaction \( A_{ij} \) can also be denoted by equation [2] with identical \( N_{i}, N_{j}, d_{ij} \) to the ones in equation [5], but with a different decay coefficient. If we use \( b_{E} \) for the decay coefficient of the EI, and \( b_{O} \) for the observed interaction, plugging the formula \( A_{ij} \) into equation [5] yields to equation [6].
From this function, it is apparent that the larger $b_E$ is than $b_O$, namely, the faster $EI$ decays with distance than $OI$, the higher the modularity. This is because modularity is a value similar to the concept of residual. Modularity greater than one signifies that the expected decay is stronger than the observed (eq. [6]). But what is the effect on the detected results? What different spatial patterns will emerge under varied relationships between $b_E$ and $b_O$?

In our study, we will compare, by hypothesis testing, not only the difference in detection results under the two basic scenarios: 1) $b_E$ is greater than $b_O$; and 2) $b_E$ is smaller than $b_O$, but also, we examine different subclasses when $b_E$ is greater or smaller than $b_O$ to different extents. As shown in Figure 1, both cases have the same OI (i.e., the same $b_O$ since the study area does not change), but case a has a larger $b_E$ (= 1.65) while case b has a smaller one (= 1.59). Though both cases are decaying at the same pace, the OI in the former case literally decays slower compared with the latter case, because it is supposed to decay faster. Therefore, it makes sense that case a yields more cells assigned to exclaves with longer distance. The following sections will illustrate the simulation and hypothesis test to justify our assumptions.

Verifying the Assumption with Simulation and Hypothesis Test

To verify our assumption, we primarily conduct a series of control experiments. In practice, the value of $b_O$ is usually drawn by fitting empirical data to the model curve, such as based on Poisson distribution [12] or algebraic simplification [13]. While in
this study, we have a direct access to $b_O$ because we simulate the OI. We require it to follow equation [3], where we subjectively determine the value of $b_O$. To ensure the robustness of this test, we set 6 values for $b_O$ instead of one, i.e., $b_O = [0.5, 0.8, 1.0, 1.2, 1.5, 2.0]$. Every time we run the test, one of the six values will be picked out and plugged into equation [4]. Given that the actual interactions do not perfectly follow a gravity model, we add a random fluctuation to each simulated OI at a specific distance $d_{ij}$; as a result, even a pair of cells with exactly the same size and the same distance as another pair will be assigned slightly different interaction strength (SI Text B). For each simulation based on a specific $b_O$, we simulate 10 times to exclude the effect of randomness on our conclusion. Therefore, we produce 10 different simulated interactions associated to a particular $b_O$.

Then we simulate the values of the distance decay coefficient for EI (i.e., $b_E$). According to our assumption, we should build the scenarios when $b_E$ is greater or smaller than $b_O$, and to different extents. We achieve this by adding on $b_O$ a variation $d$ to get $b_E$, namely,

$$b_E = b_O + d.$$ [7]

The variation $d$ is a series of 9 positive or negative values, $d = [-0.5, -0.3, -0.2, -0.1, 0, 0.1, 0.2, 0.3, 0.5]$. By cross-matching $d$ and $b_O$, we get scenarios when $b_E > b_O$ or $b_E < b_O$ (Table 1). In total there are $6*10*9 = 540$ detected results from 54 combinations of $b_E$ (i.e., $b_O + d$) and $b_O$. Among the 540 results, 240 belong to $b_E > b_O$, another 240 belong to $b_E < b_O$, and the left 60 is in $b_E = b_O$ (when $d = 0$). We do modularity-based community detection in spatial network with each combination of $b_O$ and $b_E$. To ensure that the results from different combinations are comparable, we assign cells into 100 communities, instead of assigning to the number of communities when
modularity reaches its peak value in each experiment.

<table>
<thead>
<tr>
<th>$b_E$</th>
<th>$d$</th>
<th>$b_E$</th>
<th>$d$</th>
<th>$b_E$</th>
<th>$d$</th>
<th>$b_E$</th>
<th>$d$</th>
<th>$b_E$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-0.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.2</td>
<td>-0.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.3</td>
<td>-0.2</td>
<td>0.3</td>
<td>-0.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.4</td>
<td>-0.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>-0.3</td>
<td>0.5</td>
<td>-0.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.6</td>
<td>0.1</td>
<td>0.6</td>
<td>-0.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>0.7</td>
<td>-0.1</td>
<td>0.7</td>
<td>-0.3</td>
<td>0.7</td>
<td>-0.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.8</td>
<td>0.3</td>
<td>0.8</td>
<td>0</td>
<td>0.8</td>
<td>-0.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>0.9</td>
<td>0.1</td>
<td>0.9</td>
<td>-0.3</td>
<td>0.9</td>
<td>-0.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>0.2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>-0.2</td>
<td>1</td>
<td>-0.5</td>
</tr>
<tr>
<td>1.1</td>
<td>0.3</td>
<td>1.1</td>
<td>0.1</td>
<td>1.1</td>
<td>-0.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1.2</td>
<td>0.2</td>
<td>1.2</td>
<td>0</td>
<td>1.2</td>
<td>-0.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1.3</td>
<td>0.5</td>
<td>1.3</td>
<td>0.3</td>
<td>1.3</td>
<td>0.1</td>
<td>1.3</td>
<td>-0.2</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>1.4</td>
<td>0.2</td>
<td>1.4</td>
<td>-0.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.3</td>
<td>1.5</td>
<td>0</td>
<td>1.5</td>
<td>-0.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>1.6</td>
<td>0.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.7</td>
<td>0.5</td>
</tr>
<tr>
<td>1.7</td>
<td>0.5</td>
<td>1.7</td>
<td>0.2</td>
<td>1.7</td>
<td>-0.3</td>
<td>-</td>
<td>-</td>
<td>1.8</td>
<td>0.3</td>
</tr>
<tr>
<td>1.8</td>
<td>0.3</td>
<td>1.8</td>
<td>-0.2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1.9</td>
<td>0.5</td>
</tr>
<tr>
<td>1.9</td>
<td>0.5</td>
<td>2</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2.1</td>
<td>0.1</td>
</tr>
<tr>
<td>2.2</td>
<td>0.2</td>
<td>-</td>
<td>-</td>
<td>2.3</td>
<td>0.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2.5</td>
<td>0.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Then we perform t-test to examine whether there is significant difference between any two results in terms of the amount of exclaves. To simplify the discussion, we claim two general cases (Case-1 and Case-2). Under the two cases, there are subclasses in which decays occur to different extents.

**Case-1**: \( b_E > b_O \), i.e., expected distance decay is faster than the (simulated) real decay;  
**Case-2**: \( b_E < b_O \), i.e., expected distance decay is slower than the (simulated) real decay.

The null hypothesis \( (H_0) \) proposes there is no difference in the amount of exclaves between two different combinations of \( b_E \) and \( b_O \), but we expect to reject it to support our assumption. We count the number of islands in each community and get the distribution of that variable. More than one island in a community means this community is an exclave, and more exclaves indicate a detached interaction pattern. We speculate, in Case-1, there should be significantly more exclaves than in Case-2, because in the former case interactions last farther than expected by the distance decay model while the other way around in the latter case.

We do hypothesis testing for each possible pair of the 540 detected results (including identical and different combinations) and conduct two-sample unpaired t-test based on the distribution of the amount of exclaves (eq. [8]). The distribution \( (D) \) is the absolute frequency \( (f) \) of the number of islands in a community \( (I_c) \). \( f(I_c) \) is hence quantified by the number of communities with \( I_c \) islands. When \( I_c > 1 \), the community has exclaves; otherwise, is an isolated island. For each combination of \( b_E \) and \( b_O \), we can get a distribution, so there are 540 distributions in total. We test whether the mean value of one distribution \( (D_a) \) is significantly larger or smaller than
another \( (D_b) \) (eq. [8], where \( \overline{D}_a \) and \( \overline{D}_b \) are the mean values of the first and second
distributions, \( s_a \) and \( s_b \) are the standard errors, and \( n_a, n_b \) are the sample sizes). The
null hypothesis claims no difference between \( D_a \) and \( D_b \), indicating no influence from
the value of distance decay coefficient on the pattern of exclaves. However, we hope
to reject the null hypothesis to prove the effect of the decay coefficient. Especially
when the two distributions are from Case-1 and Case-2, respectively, we would see
apparent difference in experiment results.

\[
t = \frac{\overline{D}_a - \overline{D}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}} \tag{[8]}
\]

**Results and Discussions**

**Qualitative Analysis**

We firstly illustrate the result in a qualitative way to get a general idea of the
community pattern. Figure 2 is the result of community detection with some selected
exclaves highlighted. On this map, we only show the aggregated “island” spatial
structure (i.e., a group of cells belonging to the same community and spatially
adjacent) to visualize exclaves more clearly. Comparing Figure 2-a and 2-b, it is
apparently easier to find exclaves in the former case. Exclaves emerge because
interactions happen longer than what is expected by a pre-specified gravity model for
many reasons such as business and personal social relationship. In the latter case,
however, most of the communities just contain one island, because the interactions
occur at a very local scale. Such results associates with our simulation condition that
Case-1 is $b_E > b_O$, while Case-2 is $b_E < b_O$. Such a statement will be validated by statistical testing in a quantitative manner in the following section.

Figure 2. The simulation results with different combinations of $b_E$ and $b_O$. Both maps demonstrate aggregated islands (instead of individual cells, which are smaller units than islands). a) is the result from the simulation where $b_O = 0.5$ and $b_E = 1$ (namely, $d = 0.5$). In this case, OI decays slower than EI, and therefore, urban travels
are averagely longer than the expected scenario. It thus makes sense to see one community has more islands (5 islands highlighted in the selected community) that are physically detached but functionally closely related (i.e., exclaves). b) is the result when \( b_O = 0.5 \) and \( b_E = 0.3 \) (namely, \( d = -0.2 \)). In this case, OI decays faster than EI, which is the opposite case to the above. b) shows three communities (with highlighted boundaries), each of which is composed of just 2 islands. Most of the communities do not have exclaves at all, namely, have only one island.

### Statistical Testing

With two sample unpaired t-test, we get t-scores for each pair of the 540 results. In total, there are 145,530 (i.e., \( 540^* (540 - 1) / 2 \)) results. From the t-test table (SI Text C), most of the results within the same combination of \( b_E \) and \( b_O \) (i.e., the results of the 10 simulations with the same scenario) fail to reject the null hypothesis, but a majority of the results across different combinations reject it. The combinations of parameters, therefore, make a difference in the pattern of exclaves.

Moreover, varied relationships between \( b_E \) and \( b_O \) are different in affecting exclaves. Generally, there are three types of t-scores based upon the relationship, 1) one community detection result is from \( b_E > b_O \) (Case-1) while the other from \( b_E < b_O \) (Case-2), 2) both detection results are from \( b_E > b_O \) (Case-1), and 3) both detection results are from \( b_E < b_O \) (Case-2). We find that the t-scores (SI Text C) of the first type usually yield much larger absolute values of \( t \), because the two results under comparison are from two opposite cases. But in the second or the third type, t-scores are often smaller, because the patterns are relatively similar. In each type, t-scores
vary as well. We define the gap between $b_E$ and $b_O$ as variation $d$ (eq.[7]), and the
difference between the two results’ variations is called $\Delta d$. A larger absolute value of
$\Delta d$ mostly signifies a higher $t$-score. Therefore, exclaves are generally determined by
how different $b_E$ and $b_O$ are, instead of the absolute magnitude of $b_E$ and $b_O$. The more
different $b_E$ and $b_O$, the more exclaves.

Conclusions and Indications

The distance decay coefficient is a parameter indicating the extent to which a city is
agglomerated or distributed. Stronger distance decay means that interactions are
relatively local rather than dispersed, in which case people are more likely to make
shorter trips. Note that “short” is defined relatively referring to a city’s scale instead of
a fixed distance. A pattern with more exclaves signifies either potential traffic
infrastructure issues, or implicit land use type that leads to a strong commutation
requirement. Certainly more long-distance urban-scale travels are unwanted if we aim
to reduce emission and save travel expenditure.

In this study, we quantitatively examine the influence of distance decay
coefficient on the detected communities in space-embedded networks. By discussing
the relationships between the coefficients for the expected and for the observed
interactions, we suggest that interactions with longer distance beyond the expected
(by the gravity model) will be extracted out. We concentrate on the theoretical
mechanism in this work, but the indications are helpful for future urban studies. The
results indicate that if the observed interactions decay slower than expected, namely,
more long-distance interactions take place, we would find more exclaves in the result.
We give a possible application. If, for instance, we get a distance decay coefficient
from the previous several years, but detect many exclaves (with that coefficient) based
upon the interactions of a recent network, it means the interactions in the study area
are getting longer, i.e., decaying slower. A further geographical context analysis might
yield more insights on the detected changes. This work, therefore, is meaningful in
evaluating urban dynamics by looking at the distribution of detected community of
spatial interactions (e.g., human movement, vehicle traffic, commodity flow, resource
allocation, and information communication), especially exclaves.

Acknowledgements

The authors would like to thank Dr. Yu Liu at Peking University for guidance,
discussions, and comments on this research.

References

Space-independent Communities in Spatial Networks. Proc Natl Acad Sci U S
maximization and tree clustering: Novel ways to determine effective
geographic borders. Handb Optim Complex Networks Theory Appl.


Supporting Information

A. Derivation from the Deterrence Function to the Power Law Decay

Expert el al.’s modularity model is (adopted from eq. [3] in [11]):

\[
Q = \frac{1}{2m} \sum_{C \in P_1, j \in C} |A_{ij} - P_{ij}|
\]

[S1]

where \( P_{ij} \) is the expected flow between community \( i \) and \( j \) (eq. [S2]):

\[
P_{ij} = N_iN_jf(d_{ij})
\]

[S2]

The traditional gravity model is:

\[
P_{ij} = k \frac{N_iN_j}{d_{ij}^b}
\]

[S3].

By plugging equation [S3] into the left side of equation [S2], we get

\[
f(d_{ij}) = \frac{k}{d_{ij}^b} = \frac{P_{ij}}{N_iN_j}
\]

[S4].

If we have observed flow \( A_{ij} \) between each pair of places \( i \) and \( j \), we can get the estimated value of \( b \) (namely, \( \hat{b} \)) by fitting data with equation [S4], i.e., by substituting \( P_{ij} \) with \( A_{ij} \) to get equation [S5], which is an individual level function for distance decay, since it fits the curve by each particular data plot.

\[
f(d_{ij}) = \frac{k}{d_{ij}^b} = \frac{A_{ij}}{N_iN_j}
\]

[S5]

However, in Expert et al.’s method, distance decay is approximated by
where the numerator is an aggregation of flows at a distance bin centered at \( d \) (namely, a distance band \([d - \Delta, d + \Delta]\)), and the denominator is the aggregation of the products of node sizes at the two ends of each of the flows, i.e., \( A_{ij} \). Up to here, it is obvious to see the correspondence between equation [S5] derived from traditional gravity model and equation [S6] proposed by Expert et al. [11].

**B. The Simulation of the Observed Interactions**

The *Observed Interactions* (OIs) in our context are not really observed, but instead, simulated. We call them “observed” since in the simulated cases, these interactions are supposed to be what is happening in reality. We have shown, from prior knowledge, that the real interactions practically follow a distance decay function, i.e., gravity model, but do not fit the model perfectly. In reality, there is always random fluctuation, just like the residual in a regression model. Therefore, we simulate the interactions by assigning a distance decay coefficient for the OI, and adding a random error term onto the value exactly on the model curve. The formula is equation [S7], where \( F_O(i, j) \) is the simulated OI between node (i.e., place) \( i \) and \( j \); \( k, N_i, N_j, d_{ij}, \) and \( \delta \) have the same meanings as aforementioned, and \( \delta \) is the random fluctuation. We simulate the value for each pair of \( i \) and \( j \) over the network.

\[
F_O(i, j) = k \frac{N_i N_j}{d_{ij}^{b_O}} + \delta \quad [S7]
\]

**C. \( t \)-test Result Table**

The spreadsheet named “Exclaves_ttest” in the Excel file\(^2\) demonstrates the \( t \)-test results for the distribution of the number of islands for each community detection result. From row 1 to row 14, the first column is the number of islands.

\(^2\) The spreadsheet can be downloaded at [http://www.geog.ucsb.edu/~sgao/download/BDUIC-SITable.xlsx](http://www.geog.ucsb.edu/~sgao/download/BDUIC-SITable.xlsx)
(#Islands), and the following 540 columns are the number of communities (i.e., frequency distribution) corresponding to #Islands for each detection result. For each 10 columns, the first row labels $b_O$ and $d$ in the format of “$b_Od$”. The 10 columns that share the same label are the 10 randomly simulated results with the same combination of $b_O$ and $b_E$. The 15th row is the average number of islands per community over the network, and the 16th row is the standard deviation for this variable. The 18th row displays the number of communities that have exclaves (namely, with more than one island), which is the sum of the values from row 2 to row 14. From row 21 on are the t-score for each pair of the 540*540 results (but with just the lower half triangle). For simplification, we remove the diagonal cells. Significant t-scores are highlighted with yellow. A higher absolute value of t-score indicates a more significant difference between two distributions. We find the tests with one positive $d$ and the other negative $d$ usually yield higher t-scores when $b_O$’s are the same; the larger the difference between the two $d$’s (i.e., larger $\Delta d$), the higher the score. For example, the absolute value of t-score at B27 (one with $b_O = 0.5, d = -0.1$, and the other with $b_O = 0.5, d = -0.1$) is much smaller than at B96 (one with $b_O = 0.5, d = 0.5$, and the other with $b_O = 0.5, d = -0.1$); in the former case, $|\Delta d|$ is 0, while in the latter, $|\Delta d|$ is 0.6.

The second spreadsheet named “Between comb diff” shows the results of the average number of exclaves per community between different combinations of the distance decay coefficients. In the t-test, one distribution is of the 10 simulated results with a set of coefficient combination, while the other is based upon another set of coefficients. From row 6 on are the p-values for each test. A p-value greater than 0.05 is highlighted with red color, signifying an insignificant difference. We find that the results are mostly significant except for some rare cases. Especially, the results with a larger $\Delta d$ yields a smaller p-value, which specifies a higher significance. For instance, C28 is the p-value between one result with $b_O = 1.0, d = 0.1$ and the other with $b_O = 0.5, d = -0.2$, which is much smaller than the p-value at G28 (one with $b_O = 1.0, d = 0.1$ and the other with $b_O = 0.5, d = 0.2$).
Seeing Chinese Cities through Big Data and Statistics

Jeremy S. Wu\textsuperscript{1} and Rui Zhang\textsuperscript{2}

July 10, 2014

\textsuperscript{1}Jeremy S. Wu, Ph.D., Consultant

\textsuperscript{2}Rui Zhang, Cooperation and Development Manager of Public Relations and Communications Department, Digital China Holdings Limited, Beijing, China

This research was supported in part by Digital China Holdings Limited and East China Normal University. Correspondence concerning this article should be addressed to Jeremy S. Wu, 1200 Windrock Drive, McLean Virginia 22012. We wish to thank Dr. Carson Eoyang and Dr. Xiao-Li Meng for their valuable comments, corrections, and edits.

Contact: Jeremy.S.Wu@gmail.com
Abstract

China has historically been an agricultural nation. China’s urbanization rate was reported to be 18% in 1978 when it began its economic reforms. It has now become the second largest economy in the world. Urbanization in China increased dramatically in support of this economic growth, tripling to 54% by the end of 2013. At the same time, many major urban problems also surfaced, including environmental degradation, lack of affordable housing, and traffic congestion. Economic growth will continue to be China’s central policy in the foreseeable future. Chinese cities are seriously challenged to support continuing economic growth with a high quality of life for their residents, while addressing the existing big city diseases. The term “Smart City” began to appear globally around 2008. Embracing the concept allows China to downscale its previous national approach to a more manageable city level. By the end of 2013, China has designated at least 193 locations to be smart city test sites; a national urbanization plan followed in March 2014. The direction of urban development and major challenges are identified in this paper. Some of them are global in nature, and some unique to China. The nation will undoubtedly continue to build their smarter cities in the coming years. The first integrated public information service platform was implemented for several test sites in 2013. It provides a one-stop center for millions of card-carrying residents to use a secured smart card to perform previously separate city functions and consolidate data collection. The pioneering system is real work in progress and helps to lay the foundation for building urban informatics in China. This paper also discusses the evolving research needs and data limitations, observes a smart city in progress, and makes some comparisons with the U.S. and other nations.

Keywords: Urbanization, Smart City, Statistics, Big Data, Urban Informatics, China
Seeing Chinese Cities through Big Data and Statistics

Background and Overview

China has historically been an agricultural nation. Its urbanization rate was reported to be about 11% in 1949 and 18% in 1978. Subject to differences in definition (Qiu, 2012), the U.S. urbanization rate was estimated to be at 74% in 1980 (U.S. Census Bureau, 1990).

1978 was also the year China began its economic reforms known as “Socialism with Chinese characteristics.” It introduced market principles and opened the country to foreign investment, followed by privatization of businesses and loosening of state control in the 1980s.

In the last 36 years, China leapfrogged from the ninth to the second largest economy in the world in gross domestic product (GDP), surpassing all other countries except the U.S. (Wikipedia, “List of countries by GDP (nominal)”). The poverty rate in China dropped from 85% in 1981 to 13% in 2008 (World Bank, “Poverty headcount ratio at $1.25 a day”).


Migration of rural workers to meet the urban labor needs accounted for most of the growth of the Chinese urban population to 711 million. However, under the unique Chinese household registration system known as Hukou (Wikipedia, “Hukou System”), the registered rural residents living in the city are not entitled to the government benefits of the city, such as health care, housing subsidy, education for children, job training, and unemployment insurance. Conversion from the rural to urban registration status has been practically impossible.

<table>
<thead>
<tr>
<th>Table 1. Number of Chinese Cities 1978-2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
</tr>
<tr>
<td>Cities</td>
</tr>
<tr>
<td>≥10 million (Megacity)</td>
</tr>
<tr>
<td>5-10 million (Extra Large City)</td>
</tr>
<tr>
<td>3-5 million (Extra Large City)</td>
</tr>
<tr>
<td>1-3 million (Large City)</td>
</tr>
<tr>
<td>0.5-1 million (Mid-size City)</td>
</tr>
<tr>
<td>≤0.5 million (Small City)</td>
</tr>
<tr>
<td><strong>Towns</strong></td>
</tr>
</tbody>
</table>

Source: The State Council of China (2014)
This disparity has become a major concern for social discontent in a nation of almost 1.4 billion people. Although 52.6% of the Chinese population lived in cities in 2012, only 35.3% were registered urban residents. The gap of 17.3% is known as the “floating population,” amounting to 234 million people and well exceeding the entire U.S. labor force of 156 million people. Figure 1 shows that this gap has been widening since 1978.

In addition to the social inequity caused by the Hukou system, there is an increasing geographical divide. Figure 2 shows that the eastern region of China is more densely populated than the rest of the nation (Beijing City Lab, 2014). Five out of the six megacities in 2010 are located on the east coast. These megacity clusters occupy only 2.8% of the nation’s land area, but contain 18% of the population and 36% of the GDP. While the east coast is increasingly suffocated by people and demand for resources, the central and the western regions lag behind in economic development and income.

Reliance on urban land sale and development to generate local revenue during the reform process has led to high real estate and housing prices and conflicts with the preservation of historical and cultural sites in all regions. Conversion of land from agricultural use is also raising concerns about future food supply.
At the same time, “big city diseases” surfaced and became prevalent in China, including environmental degradation, inadequate housing, traffic congestion, treatment of sewage and garbage, food security, and rising demand for energy, water and other resources. Many of these issues have been discussed domestically and internationally (e.g., Henderson, 2009; Zhang, 2010a and 2010b; United Nations Development Program, 2013).

So, where is China heading in terms of economic growth and urbanization? The answer to this question is unambiguous. The urbanization goal of 51.5% for the 12th Five-Year Plan (2011-2015) has already been exceeded (National People’s Congress, 2011).

The 2014 Report on the Work of the Chinese Government (K. Li, 2014), which is similar to the annual State of the Union in the U.S., states that “economic growth remains the key to solving all (of China’s) problems” and urbanization is “the sure route to modernization and an important basis for integrating the urban and rural structures.”

On March 16, 2014, the State Council of China (2014) released its first six-year plan on urbanization for 2014-2020. The comprehensive plan covers 8 parts, 31 chapters, and 27,000 words, providing guiding principles, priorities for development, and numerical and qualitative goals. Under this plan, China sets a goal of 60% for its urbanization by 2020.

In other words, in China’s pursuit of moderate prosperity and harmony, economic growth and urbanization will continue to be its central policy in the foreseeable future. However, more balanced efforts are pledged to manage the big city diseases and to improve the quality of life.

**Approach and Goals**

In the early days of reform, China took the trial-and-error approach of “feeling the rocks to cross the river” when infrastructure and options were lacking. Over time, the original simple economic goals were challenged by conflicting cultural and social values. More scientific evaluations are needed to minimize costly mistakes made by instinctive decisions.

After 30-plus years of reform, Chinese President Xi Jinping (2014) acknowledged that “…the easier reforms that could make everyone happy – have already been completed. The tasty meat has been eaten up. The rest are tough bones to crack.” In other words, difficult choices will have to be made in China’s reform process.

Chinese Premier Li Keqiang (2014) pledged to carry out “people-centered” urbanization in his 2014 work report and cited three priorities on three groups of 100 million people each:
- Granting official urban Hukou status to 100 million rural people who have already moved to cities;
- Rebuilding rundown, shanty city areas and villages inside cities where 100 million people currently live (Figure 3 shows an image of contrasting buildings in Shanghai, the largest city in China);
- Guiding the urbanization of 100 million rural residents of the central and western regions into cities in their regions.

Table 2 reproduces the 18 key numerical goals for 2020 along with the 2012 benchmarks under the national urbanization plan.

There are now two key goals in regard to the urban population: raising the level of residents living in cities to 60% and the level of registered urban residents to 45%, thereby reducing the floating population from the current 17.3% to 15% in six years. The other key goals promote the assimilation of migrant rural workers into city life, improving urban public service and quality of life, and protecting land use and the environment.

There are less specific qualitative goals in the national urbanization plan. For example, Chapter 17 mandates “Three Districts and Four Lines” in each city. The three districts are defined as areas forbidden from, restricted from, and suitable for construction respectively. Four types of zones will be drawn by color lines: green line for ecological land control; blue line for protection of water resources and swamps; purple line for preservation of historical and cultural sites; and yellow line for urban planning and development.

How these districts and zones will be created and sustained has not been specified.

China is pressing forward with concurrent modernization in agriculture, industrialization, information technology, and urbanization. Under the urbanization plan, the central government is responsible for strategic planning and guidance. Authority is delegated to the provincial and municipal levels through political reform. Local administrators are encouraged to innovate,
Conversion from rural to urban Hukou registration is now officially allowed and encouraged, but the process will be defined by the individual cities, under the general rule that the conversion will be more restrictive as the population of the city increases.
Implementation of the urbanization plan will be complex and challenging. “Feeling the rocks to cross the river” is no longer adequate as a stand-alone approach. There is recognition that today’s urban development requires a proactive, data-driven strategy that must be efficient and innovative with intelligent use of the latest technologies and information.

**Emerging Role of Statistics and Technology**

The national urbanization plan provides an unprecedented opportunity for statistics and technology to support and monitor the implementation of policies in China. Chapter 31 prescribes the role of defining metrics, standards, and methods to establish a sound statistical system, monitoring the activities dynamically, and performing longitudinal analysis and continuing assessment of the progress of urbanization according to the development trends.

The specification of dynamic monitoring and longitudinal analysis reflects advanced thinking, compared to the current static, cross-sectional reports. Yet how the statistical monitoring system will be implemented also remains unclear at this stage.

There are many ways to cross a river. “Feeling the rock” is a popular Chinese euphemism for taking a trial-and-error approach when the Chinese economic reform began in 1978. While it is still an important part of scientific discovery, total reliance on trial and error can be inefficient, costly, and even dangerous for the governance of a nation or a city.

More preferable is to measure the depth of the river at different times, collect relevant and reliable data, analyze the results scientifically to search for ideal crossing points, evaluate the pros and cons of the options, and make informed, intelligent decisions so that many can cross the river fairly and safely.

Many developed nations have been using this data-driven approach to manage knowledge for their businesses (e.g., Sain and Wilde, 2014). It is the assumed approach that China will also take for governance in this paper. Figure 4 shows a Data-Information-Knowledge-Wisdom (DIKW) hierarchy model for this process.

![Figure 4. The DIKW Hierarchy Model](image)
The foundation of scientific knowledge and wisdom is to observe facts and collect data. However, data in their raw form have little or no meaning by themselves. Not all data have adequate information value or are useful for effective decision making.

Statistics, both as a branch of science for knowledge discovery and a set of measurements, provides context and value by converting useful data into relevant information. Knowledge is gained and accumulated from information, and used as the basis for making wise decisions. The decisions will not be correct all the time, but the scientific process promotes efficiency and minimizes errors, especially when conducted with integrity, objectivity, and continuous improvement.

Technology is not explicitly shown in the DIKW model because the concept was practiced well before modern information technology began to emerge in the 1950s. Today the base of the pyramid is greatly expanded by technology, and the transformation of data into information is accelerated. However, the process is also contaminated by hype, useless data, and misinformation (Harford, 2014; Wu, 2014).

a. Traditional Statistics and Big Data

Census data have been used for governance of nations for centuries. A census is comprehensive, but data collection is costly and time consuming for producing static results at a point in time. Random surveys were later introduced based on probability theory to produce scientifically reliable information with proper design and a relatively small amount of data.

Together censuses and random surveys form the statistical foundation based on structured data (Webopedia, “Structured data”) in the 20th century. Developed nations have used them effectively for policy and decision making, with design and purpose, over the past hundred years.

At the turn of this century, massive amounts of data began to appear in or were converted from analog to digital form, allowing direct machine processing (Hilbert and Lopez, 2012). At the same time, huge storage capacity and computing power became available at reasonable cost. It is defined in this paper as the beginning of the Big Data era. Big Data was not a well-known term in China until Tu (2012) published the first Chinese-language book on the topic.

Unstructured data (Wikipedia, “Unstructured data”) are typically created by sensors, social media, e-Commerce, and automated sources. Here the meaning of data may expand beyond numbers to include text, map, image, sound, and multimedia. Their sheer volume and
dynamic nature provide enormous possibilities for application. On the other hand, incomplete and unstructured data are difficult to process and may even be meaningless to analyze in the absence of design and purpose. Although data mining is commonly mentioned as a promising approach to extract information for commercial purposes, their reliability and value can be suspect, especially for the purpose of governance (e.g., Marcus and Davis, 2014; Lazer, Kennedy, King, and Vespignani, 2014). Few of the key numerical goals in the national urbanization plan can be measured meaningfully or reliably by unstructured data.

Integration of structured data derived from administrative records to create longitudinal data systems was the first realized benefit of Big Data for government statistics.

Under the Longitudinal Employer-Household Dynamics (LEHD) program, the U.S. Census Bureau merges unemployment insurance data, social security records, tax filings and other data sources with census and survey data to create a longitudinal frame of jobs. It is designed to track every worker and every business in the nation dynamically through the relationship of a job connecting them, with data updated every quarter. The comprehensive longitudinal summaries protect confidentiality. They provide insights about patterns and transitions over time, which are not available from the traditional cross-sectional statistics.

Similar efforts to build longitudinal data systems for education (Data Quality Campaign, n.d.) and health care (Wikipedia, “Health Information Technology for Economic and Clinical Health Act”) are underway in the U.S. The 2020 U.S. census will also be supplemented by the integration of administrative records (Morello, 2014).

b. National Basic Data Systems and Identification Codes

In principle, a nation is composed of its people, businesses, government, and environment, which in turn form an economy. More than a decade ago, the State Council of China (2002) issued guidance to create four National Basic Data Systems as part of e-Government - longitudinal frames of people, enterprises, and environment/geography respectively with the fourth system as an integration of the first three to form a unified macroeconomic data system.

These nationwide data systems possess the desired characteristics of a 21st century statistical system (Groves, 2012; Wu, 2012; Wu and Guo, 2013). They help to transition the Chinese government’s role from central control to service for citizens and to establish a foundation for data sharing and one-stop integrated service nationwide.
Heavy investment followed to establish and implement definitions, identification codes, standards, and related infrastructure.

Identification codes are the keys to unlocking the enormous power in Big Data (Wu and Ding, 2013). A well-designed code matches and merges electronic records, offers protection of identity, provides basic description and classification, performs initial quality check, and facilitates the creation of dynamic frames.

As early as 1984, China began to build an infrastructure with its citizen identification system (Wikipedia, “Resident Card System”). A sample Chinese citizen card (Figure 5) displays the citizen identification code, name, gender, ethnicity, birthdate, address, issuing agency, dates of issuance and expiration, and a photograph.

The 18-digit citizen identification code, introduced in 1999, includes a Hukou address code, birthdate, gender, and a check digit. It is issued and administered by the Ministry of Public Security. The citizen code is uniquely and permanently assigned to the cardholder. The card is capable of storing biometric information. It is increasingly required for multiple purposes, such as the purchase of a train ticket for travel.

The U.S. does not have a comparable national citizen card system. Recent renewed discussions about adding an image of the cardholder to the Social Security card was met again with controversy (e.g., Bream, 2014; Eilperin and Tumulty, 2014).

China has also established the system of National Organization Codes under the National Administration for Code Allocation to Organizations. The 9-digit organization code includes a check digit; it is a unique identification and linking variable to store and retrieve information about companies, institutions, communities, government agencies and other registered organizations in China, functioning like the Employer Identification Number in the U.S. (Wikipedia, “Employer Identification Number”).

Figure 5. Sample Chinese Citizen Identification Card

China has laid a sound foundation for building dynamic frames through these initiatives. However, by 2008, the national approach to create Basic Data Systems was becoming too complex with too many structural, legal, and practical obstacles to overcome.

Shen (2008) reported that the Basic Data System on environment and geography was essentially complete but lacked real application. The Basic Data System on population was burdened by the inclusion of over 100 variables, each with a different degree of sensitivity. The Basic Data System on enterprises faced the strongest resistance to data sharing by various agencies with overlapping responsibilities. The Basic Data System on macroeconomics was stalled without the first three data systems in place.

The LEHD program in the U.S. went through comparable experiences. The national approach faced resistance to data sharing so that the approach had to be strategically adjusted to the state level before data can be re-assembled to the national level.

The Basic Data Systems were relegated to long-term development until the release of the national urbanization plan in 2014. Mandates are now revived and issued for their accelerated development and implementation. For example, the Basic Data System on population is expected to link to cross-agency and cross-regional information systems for employment, education, income, social security, housing, credit services, family planning, and taxation by 2020. The citizen identification code is also mandated to be the only legal standard for recording, inquiring, and measuring population characteristics in China the same year.

c. The Rise of Smart City

The term “Smart City” began to appear globally around 2008 as an extension to previous development of e-Government and digital cities. In general, a city is considered “smart” (Wikipedia, “Smart City”) when “investments in human and social capital and traditional (transport) and modern communications infrastructure fuel sustainable economic development and a high quality of life, with a wise management of natural resources, through participatory action and engagement.” Data collection, processing, integration, analysis and application are at the core of constructing smart cities.

In practical terms, embracing the concept of smart city will allow China to downscale the original national approach to the more manageable city level, while protecting past investments and permitting aggregation to the provincial or regional level.
Table 3 describes the direction to develop smart cities as outlined in the national urbanization plan. At the end of 2013, the Chinese Ministry of Housing and Urban-Rural Development has designated 193 locations to be smart city test sites (baidu.com, “National Smart City Test Sites”). They are expected to undergo 3-5 years of experimental development.

The Chinese Ministry of Science and Technology has also named 20 smart city test sites (Xinhuanet.com, 2013). They are expected to spend three years to develop templates of cloud computing, mobile networks, and related technologies for broad implementation.

<table>
<thead>
<tr>
<th>Table 3. Direction of Smart City Development</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>01 Broadband Information Network</strong></td>
</tr>
<tr>
<td>Replace copper by fiber-optics. Implement fiber-optic network covering practically all urban families at connection speed of 50Mbps, 50% families reach 100Mbps, and some families reach 1Gbps in well-developed cities. Develop 4G network and accelerate public hot spots and WiFi coverage.</td>
</tr>
<tr>
<td><strong>02 Information Technology for Planning and Management</strong></td>
</tr>
<tr>
<td>Develop digital city management, promote platform development and expand functions, establish a unified city geospatial information platform and building (structure) database, build public information platform, coordinate the digitization and refinement of urban planning, land use, distribution network, landscaping, environmental protection and other municipal infrastructure management</td>
</tr>
<tr>
<td><strong>03 Intelligent Infrastructure</strong></td>
</tr>
<tr>
<td>Develop intelligent transportation to guide traffic, command and control, manage adjustments and emergencies. Develop smart grid to support distributed access to energy and intelligent use of electricity by residents and businesses. Develop intelligent water services to cover the entire process from quality and safety of supply to drainage and sewage. Develop intelligent information network to manage urban underground space and pipes. Develop intelligent buildings to manage facilities, equipment, energy consumption, and security</td>
</tr>
<tr>
<td><strong>04 Public Service Streamlining</strong></td>
</tr>
<tr>
<td>Establish cross-agency, cross-regional business collaboration, sharing of public service information service system. Use of information technology and innovation to develop urban education, employment, social security, pension, medical and cultural service model.</td>
</tr>
<tr>
<td><strong>05 Modernization of Industrial Development</strong></td>
</tr>
<tr>
<td>Accelerate the transformation of traditional industries, promote use of information technology, digitization, and networking to transition to intelligent service models for manufacturing. Actively develop and integrate information services, e-commerce and logistics to nurture innovation and new formats</td>
</tr>
<tr>
<td><strong>06 Refinement of Social Governance</strong></td>
</tr>
<tr>
<td>Strengthen the application of information to monitor market regulations and the environment, credit services, emergency protection, crime prevention and control, public safety and other areas of governance, establish and improve relevant information service system, innovate to create new model of social governance</td>
</tr>
</tbody>
</table>

Source: The State Council of China (2014)

**Current State**

The issuance of a City Resident Card is a concrete first step of implementation for aspiring smart cities to provide one-stop service and consolidate data collection.

The multi-functional card may be used for social security or medical insurance purposes, as well as a debit card for banking and small purchases. Depending on the city, the City
Resident Card may also be used for transportation, public library, bicycle rentals, and other governmental and commercial functions yet to be developed.

During the application process, the citizen identification code is collected along with identification codes for social security and medical insurance, residence address, demographic data, and family contact information, facilitating linkage to other data systems and records.

The current smart resident cards in use in China (Figure 6) vary from city to city, but they typically contain two chips and a magnetic memory strip.

Digital China (2013), a major technology service provider in China, implemented the first integrated public information service platform for several Chinese smart city test sites in October 2013. It provides a one-stop center as an additional channel of service for millions of card-carrying residents, who can use the secured smart card to perform previously separate functions.

The city of Wuhan announced its “Wuhan Big Data Industrial Development Action Plan (2014-2018)” in April (Smarterchina.com.cn, 2014). It includes the establishment of seven cloud computing centers on governance, geo-spatial information, data management, education, multimedia, quality monitoring, and automobile network respectively.

On June 5, World Environment Day, the city of Dongguan announced the release of real-time data on four types of emission from its incineration plants in a publicly accessible website (cn-hw.net, 2014). It is the first of its kind in China.

**Urban Informatics**

The above pioneering activities are modest, but they are real work in progress and will lay the foundation for urban informatics in China. They represent the very early results of
China’s total investment into smart city development, which is estimated to exceed ¥2 trillion ($322 billion) by 2025 (Yuan, 2014).

Urban informatics, meaning the scientific use of data and technology to study the status, needs, challenges, and opportunities for cities, is presently not a well-known concept in China. It covers both unstructured and structured data, collected with and without design or purpose. The defining characteristics of urban informatics will be the sophisticated application of massive longitudinal data, integration of multiple data sources, and rapid and simple delivery of results, while strictly protecting confidentiality and data security and assuring accuracy and reliability.

China’s urbanization has high risks and high potential returns. Its enormous scale implies potential inefficiency and waste. Although experimentation is encouraged, urban informatics must crawl, walk, and run at the same time. The urgency leaves little time and room for research and development.

There are many challenges and needs in establishing urban informatics as a mature field of study in China. We select four major topics for discussion.

**a. Need for Change in Culture**

There is no assurance that internal resistance to data sharing and standards can be overcome in China despite mandates, political reforms, downscaling, and cloud computing (e.g., UPnews.cn, 2014). A major risk of a de-centralized approach is the formation of incompatible “information silos” such that the systems cannot inter-operate within or between cities.

This challenge is not unique to China. The U.S. had more than 7,000 data centers in the federal government alone in 2013; about 6,000 of them were considered “noncore.” Many of them do not communicate with each other and are costly to maintain. Although a major consolidation initiative was started by the White House in 2010, progress has been slow (CIO.Gov, 2014; Konkel, 2014).

However, open data-based governance and research are relatively new concepts in China. Although their value is recognized and advocated in the central plans, Chinese officials are not well known for their support of open data policy and data sharing, or their full awareness of modern statistical or environmental issues.

The principles of statistical quality control and management were first proposed by Walter A. Shewhart (Wikipedia, “Quality Management”) in 1924. Emphasizing the use of
statistical methods and “profound knowledge,” Deming (1994) made significant contributions to Japan’s post-war “economic miracle” and later the quality management movement in the U.S. The International Organization for Standardization 9000 series of standards are perhaps the best known product on quality management today from the non-government organization made up of members from 162 countries including China. While these statistical principles and thinking originated in the context of industrial production, they are equally applicable to governance.

The National Bureau of Statistics of China relies heavily on data supplied by provincial and local governments. Intervention and data falsification by local authorities are occasionally reported in China (e.g., Wang, 2013), including the famed GDP. For example, the incomplete 2013 GDP of 28 out of 31 provinces and cities already exceeded the preliminary 2013 total national GDP by ¥2 trillion or 3.6% (e.g., D. Li, 2014). Credibility and public confidence in China’s statistics are not high.

Tu (2014) made an exceptional observation that China has not yet developed a culture of understanding and respect for data. It is in sharp contrast with how the city of Helsinki in Finland promotes open data with cost-sharing and regionalism (Sulopuisto, 2014). Changing this culture is a challenge without historical precedent in China.

b. Need for Statistical Thinking and Design

The Chinese statistical infrastructure is recent and fragile. The first Chinese decennial census on population began in 1990 while the U.S. started 200 years earlier; the first Chinese consolidated economic census was conducted only 10 years ago in 2004. Random surveys seldom include detailed documentation on methodology.

Requiring dynamic monitoring and longitudinal analysis by the Chinese government is refreshing in the national urbanization plan. Its implementation faces many statistical and technological issues, including record linkage and integration, treatment of missing or erroneous data, ensuring data quality and integrity, retrieval and extraction of data, scope of inference, and rapid delivery of results. Some of the terms in use, such as “talented persons,” “green buildings,” and “information level,” do not have commonly accepted definition or standard meaning.

In the collection of data about a person, some of the characteristics such as gender and ethnicity remain constant over time; some change infrequently or in a predictable manner such as age, Hukou, and family status; some change more frequently such as education level, income
level, employment, and locations of home and work; and others change rapidly such as nutritional intake, use of water and electricity, or opinion about service rendered. Measurement of these characteristics must be made with appropriate frequency, completeness, and quality so that reliable data can be collected to easily and rapidly describe and infer about the population. The definitions must be consistently applied across locations and time so that the results can be compared and meaningful temporal or spatial patterns can be discovered and studied. The base unit may extend from a person to a family or a household. These considerations also apply to an enterprise or a defined geo-space.

Not all available Big Data are relevant for governance; but Big Data include the Basic Data Systems in China. Integrated, structured data contain substantially richer information than unstructured data. Extraction of information can be optimized if it is designed and built into the top-level of the data ecosystem.

Nie, Jiang and Yang (2012) reported the disastrous consequences of mismatched records, outlying data, large variations, and unclear definitions in a Chinese national longitudinal data system on enterprises. Without proper statistical design and implementation of quality control, the data system does not support credible analysis or reliable conclusions despite the high cost of its creation and maintenance. There are few discussions of statistical design or need for quality control of data systems in China. In general, large-scale longitudinal data systems or reliable longitudinal analysis are currently lacking.

There are recent calls for exploring and understanding “Scientific Big Data” (Qi, 2014; Jiang, 2014). It is a promising sign that China may be prepared to make better use of data in scientific disciplines, in addition to the current commercial and marketing environments.

c. Need for Integration of Technology and Statistics

Yuan (2014) quoted the research firm IDC that “roughly 70 percent of government investments went to hardware installation in China, way higher than the global average of 16 percent.” While China may be strong at hardware, service and software tend to lag behind. Technology and statistics are in fact disconnected.

Figure 7 shows a mature conceptual architecture for rapid delivery of business intelligence (BI) to both casual and power users, who may be government officials, academic researchers, the media, or the general public.
Rapid information delivery depends critically on the soundness of the statistical design such that the underlying data are representative, quality-assured, and properly warehoused for easy extraction, transformation, and loading (ETL). It facilitates dynamic visualization and longitudinal reporting of the status and progress on the urbanization plan, as well as system performance and customer satisfaction.

Online services based on smart resident card and one-stop center have already yielded some welcomed relief to labor-intensive administrative functions and reduction of some long waiting queues, but current static monitoring reports are not connected to data collected from online services in concept or in operation. Although statistical yearbooks are beginning to appear online, interactive queries and dynamic visualization similar to the American Factfinder (U.S. Census Bureau, n.d.) are not yet available. Intelligent mapping applications similar to OnTheMap (Wu and Graham, 2009; U.S. Census Bureau, n.d.) have also not been introduced to deliver custom maps and statistical reports based on the most recent data in real time.

d. Need for Statistical Innovation

Statistics is the scientific study of data, big or small. Although it has a long history, it is not fully developed in some areas. The common belief that a large amount of data in terms of file size will be able to provide reliable inference about a population is misguided; prevailing statistical theories do not support analyses whose data are not collected according to probabilistic design. When their representation of the population is inadequate, analyzing non-randomly collected data can result in misleading conclusions.

Meng (2014) proposed that three types of statistical inference require “an expanded paradigm with greater qualitative consistency and relative optimality.” Drawing conclusions from the Basic Data Systems, created in multiple phases with data from various linked sources,
is an example of multi-phase and multi-source inference. Being able to provide estimates at user-selected geographic levels, OnTheMap exemplifies multi-resolution inference.

Discussions of statistical methods to preserve confidentiality started as early as 1993 (e.g., Rubin 1993), and was implemented in the OnTheMap application. However, the full potential of synthetic data as a method of protecting confidentiality remains to be explored and validated.

**Zhangjiagang – A Developing Chinese Smart City**

Zhangjiagang is a port city located along the Yangtze River in eastern China under the administration of Suzhou City, Jiangsu Province (Figure 8). It has an area of 999 square kilometers and a total population of 1.5 million urban and rural residents. Given its excellent harbor and richness in ore reserves, Zhangjiagang was strong in its mining, iron and steel industry, and machinery manufacturing sector, placing in the front of the top 100 counties in China. It was changed from a county to a county-level city in 1986.

![Figure 8. Zhangjiagang, Suzhou City, Jiangsu Province](image)

Source: Google Earth

The city received the UN-Habitat Scroll of Honor Award in 2008 (UN-Habitat, 2008), becoming the first county-level city to receive such an award in China. According to the Suzhou
city government (2014), the 2013 GDP for Zhangjiagang city was about ¥214.5 billion, an increase of 6% over the previous year. Urban resident per capita disposable income was ¥43,400; rural per capita net income was ¥21,700, an increase of 9.3% and 11.4% respectively.

Table 4. Social/Economic Performance Targets for Zhangjiagang under the 12th 5-Year Plan

<table>
<thead>
<tr>
<th>Performance Indicator</th>
<th>2010 Actual</th>
<th>2015 Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic Development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 Gross Domestic Production (¥100M)</td>
<td>1600</td>
<td>3200</td>
</tr>
<tr>
<td>2 General Budget Revenue (¥100M)</td>
<td>116.06</td>
<td>250</td>
</tr>
<tr>
<td>3 Service Share of GDP (%)</td>
<td>38</td>
<td>45</td>
</tr>
<tr>
<td>4 New Industry Value as Share of Above-Scale Industrial Value (%)</td>
<td>21.8</td>
<td>50</td>
</tr>
<tr>
<td>5 Actual Use of Foreign Investment ($100M)</td>
<td>10</td>
<td>60 (5 year)</td>
</tr>
<tr>
<td>6 Total Imports and Exports ($100M)</td>
<td>250</td>
<td>500</td>
</tr>
<tr>
<td>7 Total Port Throughput (M tons)</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>8 Port Container Throughput (10K Twenty-foot Equivalent Units)</td>
<td>112</td>
<td>220</td>
</tr>
<tr>
<td>Innovation Capacity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 Research and Development As Share of GDP (%)</td>
<td>2.15</td>
<td>&gt;3.0</td>
</tr>
<tr>
<td>10 High-Tech Industry as Share of Above-Scale Industrial Value (%)</td>
<td>32</td>
<td>40</td>
</tr>
<tr>
<td>11 Large/Medium Size Enterprises as Share of R&amp;D Institutions (%)</td>
<td>65</td>
<td>90</td>
</tr>
<tr>
<td>12 Gross Enrollment of Higher Education (%)</td>
<td>65</td>
<td>70</td>
</tr>
<tr>
<td>13 Average Education of New Labor (Year/Person)</td>
<td>15.3</td>
<td>16</td>
</tr>
<tr>
<td>14 Talented Persons per 10K Persons (Person)</td>
<td>1380</td>
<td>2200</td>
</tr>
<tr>
<td>15 Invention Patents (Piece)</td>
<td>95</td>
<td>300</td>
</tr>
<tr>
<td>16 Science and Technology Contribution Rate (%)</td>
<td>55</td>
<td>60</td>
</tr>
<tr>
<td>Social Improvement</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17 Urbanization Rate (%)</td>
<td>63</td>
<td>70</td>
</tr>
<tr>
<td>18 Social Security Coverage (%)</td>
<td>Reach 100</td>
<td>All Covered</td>
</tr>
<tr>
<td>19 Gini Index</td>
<td>0.4</td>
<td>&lt;0.38</td>
</tr>
<tr>
<td>20 Health Workers per 1K Persons (Person)</td>
<td>6.38</td>
<td>7.2</td>
</tr>
<tr>
<td>21 Public Cultural Facilities Area per 1M People (sq. m.)</td>
<td>1400</td>
<td>1600</td>
</tr>
<tr>
<td>22 Public Stadium Area per 10K People (sq. m.)</td>
<td>28000</td>
<td>32000</td>
</tr>
<tr>
<td>Quality of Living</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23 Average Urban Resident per Capita Disposable Income (¥)</td>
<td>30400</td>
<td>53000</td>
</tr>
<tr>
<td>24 Average Rural Resident per Capita Net Income (¥)</td>
<td>14400</td>
<td>26500</td>
</tr>
<tr>
<td>25 Urban and Rural Resident Health Index (%)</td>
<td>96.44</td>
<td>&gt;97</td>
</tr>
<tr>
<td>26 Beds in Elderly Homes per 1K Elderly Residents (bed)</td>
<td>19.5</td>
<td>35</td>
</tr>
<tr>
<td>27 Social Satisfaction of Public Security (%)</td>
<td>98</td>
<td>&gt;98</td>
</tr>
<tr>
<td>28 Information Level in Daily Life (%)</td>
<td>80</td>
<td>≥90</td>
</tr>
<tr>
<td>29 Engel's Coefficient (%)</td>
<td>15.5</td>
<td>25</td>
</tr>
<tr>
<td>30 Environmental Quality Composite Index (point)</td>
<td>98</td>
<td>&gt;96</td>
</tr>
<tr>
<td>Ecological Environment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31 Garbage Harmless Treatment Rate (%)</td>
<td>75</td>
<td>&gt;95</td>
</tr>
<tr>
<td>32 Per Capita Green Area in Developed Area (sq. m.)</td>
<td>12.98</td>
<td>13.2</td>
</tr>
<tr>
<td>33 Water Consumption Reduction in ¥10K GDP vs 11th 5-Year Plan (%)</td>
<td>--</td>
<td>10</td>
</tr>
<tr>
<td>34 Energy Consumption Reduction in ¥10K GDP vs 11th 5-Year Plan (%)</td>
<td>--</td>
<td>Assigned</td>
</tr>
<tr>
<td>35 SO2, NOX Emission Reduction vs 11th 5-Year Plan (%)</td>
<td>--</td>
<td>Targets</td>
</tr>
<tr>
<td>36 COD, Ammonia Emission Reduction vs 11th 5-Year Plan (%)</td>
<td>--</td>
<td>Completed</td>
</tr>
</tbody>
</table>

Notes: ¥ = Chinese yuan; $=U.S. dollars; K=Thousand; M=Million
$1 ≈ ¥6.2

Source: Zhangjiagang City Government; Digital China
The urbanization rate for Zhangjiagang was 63% in 2010, already higher than the current average in China, exerting high pressure on its city administrators to manage its population, environment, and economic development.

Table 4 shows the performance indicators and goals for Zhangjiagang’s 12th five-year plan (2011-2015). Its economic output and port throughput are expected to double in the 5-year period. Zhangjiagang set up "new capabilities" and quantified research and development (R&D) in order to realize these goals. As in many other Chinese cities, science and technology is considered an important pillar supporting continued urbanization.

Chapter 8 of the Zhangjiagang 12th five-year plan also emphasizes the support of information technology for e-Government by raising the level of government applications; accelerating the construction of the Basic Data Systems; focusing on public and government data sharing and exchange systems; and further improving the level of inter-departmental applications.

One-stop Platform

The Zhangjiagang public website was launched in October 2013 with the above goals in mind. The front page of the website (Figure 9) contains 3 channels - My Service, My Voice, and My Space. My Service provides government and public services; My Voice connects the government and the resident through an online survey and microblogging; My Space contains the user's "digital life footprint" such as personal information and record of use.

The public website combines online and offline services through the use of the smart resident card, desktop and mobile devices, and government and community service centers, offering 621 types of services by 31 collaborating government and community organizations.

The services vary by type of access device. Desktop computers offer the most comprehensive services, including queries, more than 240 online applications, and over 130
online transactions. Mobile device users may check on the progress of their applications, using General Packet Radio Service (GPRS) positioning and speech recognition technology to obtain 56 types of efficiency services such as travel and transportation.

Figure 10 is a sample My Service page, showing the location and availability of rental bicycles in the city.

The one-stop service platform attempts to provide a unified, people-centric, complete portal, eliminating the infighting of various government agencies to build their own websites and service stations and consolidating separate developments such as smart transportation and smart health care.

The website combines a variety of existing and future smart city proposals and services. Developers will be able to link to the platform to provide their services with lower operating costs. The city government wants to have a platform to showcase information technology, introduce business services, and assist economic development, especially in e-Commerce.

The Zhangjiagang public website is designed to be an open platform for progressive development. All the applications will be dynamically loaded and flexible to expand or contract. Existing services will be continuously improved, and new functionalities added. It aims to improve public satisfaction of government service and broaden agency participation to facilitate future data sharing and data mining.

Participation of the residents in the platform will determine whether its goals will be achieved or not. In the 6-month period since its launch in October 2013, there have been 15,518 total users through real-name system certification and online registration, 31,956 visitors, and 198,227 page views. The average visit time was 11 minutes and 7 seconds. Among all the users, real-name registrants accounted for 67%, and mobile end users accounted for 44%.
Online booking of sports venues, event tickets, and long-distance travel are the most popular services to date. They show the value of convenience to the residents, who had to make personal visits in the past.

Although there is no current example of data sharing between government departments, the public website is beginning to integrate information for its residents. A user can view his/her records in a secured My Space (Figure 11).

It is already possible in the Zhangjiagang platform to create a consolidated bill of natural gas, water, electricity, and other living expenses to provide a simple analysis of household spending. Although this is elementary data analysis, it foretells the delivery of more precise future services as online activities and records expand and accumulate over time.

**Summary**

China is in the early stage of its six-year national urbanization plan, extending its economic development to also address rising social and environmental concerns. There is a defined role for statistics and urban informatics to establish norms and conduct dynamic monitoring and longitudinal analysis. Small steps have been taken to begin data consolidation in some smart city test sites, and modest progress is beginning to appear. In the next six years, cultural changes towards an objective data-driven approach, integration of statistical design and thinking into the data systems, and innovative statistical theories and methods to fully deploy meaningful Big Data will be needed to grow urban informatics in China and to achieve balanced success in its urbanization efforts. China will undoubtedly continue to advance towards building her smarter cities with Chinese characteristics, and we will be able to see more of the Chinese cities through statistics and Big Data.
References


Time-geographic relationships between vector fields of activity patterns and transport systems

Xintao Liu, Wai Yeung Yan, Joseph Y.J. Chow*

Department of Civil Engineering, Ryerson University, Toronto, ON, Canada

*Corresponding author’s email: joseph.chow@ryerson.ca

ABSTRACT

The rise of urban Big Data has made it possible to use demand data at an operational level, which is necessary to directly measure the economic welfare of operational strategies and events. GIS is the primary visualization tool in this regard, but most current methods are based on scalar objects that lack directionality and rate of change—key attributes of travel. The few studies that do consider vector-based GIS have largely looked at vector fields for individuals, not populations. A population-based vector field is proposed for visualizing time-geographic demand. The field is estimated using a vector kernel density generated from observed trajectories of a sample population. By representing transport systems as vector fields that share the same time-space domain, demand can be projected onto the systems to visualize relationships between them. This visualization tool offers a powerful approach to visually correlate changes in the systems with changes in demand, as demonstrated in a case study of the Greater Toronto Area using data from the 2006 and 2011 Transportation Tomorrow Surveys. As a result, it is now possible to measure in real time the effects of disasters on the economic welfare of a population, or quantify the effects of operational strategies and designs on the behavioural activity patterns of the population.

Keywords:
GIS, Big Data, activity travel patterns, kernel density, vector density, transport systems

In Proceedings of The Big Data and Urban Informatics Workshop 2014 at University of Illinois, Chicago
1. INTRODUCTION

Recent advances in Big Data ubiquity (see LaValle et al., 2011) have resulted in many new opportunities in urban transportation operations and planning, particularly in understanding how people travel (Schönfelder et al., 2002). Understanding travel behaviour is important for urban policymakers because sustainable cities depend on healthy mobility patterns, which in turn depend on a host of complex factors. In the past, lack of abundant data sources of travel patterns confined analysts to rely on periodic travel surveys conducted every few years (see Stopher and Greaves, 2007), which limited the use of travel data primarily to planning purposes. As a consequence, traffic and transit operations conducted on a day-to-day basis largely relied on only trip data, which led to disconnects between operations and long term planning. For one, operational measures that depend only on trip data (travel times, volumes, delay at particular corridors, etc.) provide policy-makers with information on trip characteristics, but not on the underlying economic demand from which travel is derived (Pinjari and Bhat, 2011). As a result, operational decisions often cannot be related directly to travel demand patterns, but only to trip characteristics: for example, the cost of a hurricane or a planned event quantified in terms of trip measures falls short as only proxies of direct measures of economic welfare. This lack of direct connection between travel demand patterns and operational factors due to lack of data is one reason why policymakers often cannot get a clear picture of trade-offs between operational alternatives with respect to mobility and economic efficiency. With the rise of Big Data, there is an opportunity to provide a much clearer and data-driven (instead of model-driven) connection between the two for policymakers because travel patterns can now be observed from many different sources at a day-to-day operational level.

However, the opportunity cannot yet be realized because there is no adequate methodology to visualize and quantify the travel pattern data observed on a day-to-day basis. In other words, Big Data is available for us to form “scatter plots” of travel demand patterns (as more direct indicators of economic activity) against operational factors, but we still lack the “scatter plot” device. Visualization and measurement of spatial-temporal patterns can be done using geographical information systems (GIS), and time geography is the specific field of visualizing and measuring societal spatial-temporal activity patterns throughout the day (Hägerstrand, 1970). However, there is no purely data-driven means of evaluating population-level time-geographic effects due to changes in a transportation system. For example, policymakers currently have no tool to adequately visualize and quantify the real-time effects that a major blizzard or flood has on a population’s travel patterns and subsequent economic welfare, even if that data were available. There are two reasons for this gap.

The first reason is related to the lack of consideration for travel disutility in GIS methods used for time geography. Individuals’ activity patterns are often captured using activity trajectories under the framework of 3D time-geography where time of day is a third dimension. However, the shapes in those spaces are still only static lines (and prisms) that do not include direction or rate of change. Since travel is inherently a rate of change with directionality, impacts measured using changes in static points or lines do not provide a full picture, as argued by Miller and Bridwell (2009). As a result, more recent developments in GIS have considered vector fields, where each point of a space is associated with a vector that includes both magnitude and direction. Miller and Bridwell (2009) discuss “anisotropic cost fields” as direction-specific costs—the closest to addressing this first gap—but it leads to the second major gap: the vector fields have only been used to construct space-time prisms for individuals.
We propose a novel GIS methodology for time geography, under a ubiquitous data setting where travel trajectories can be continuously collected from a sample population. Miller and Bridwell’s (2009) theory of anisotropic cost fields for individuals is extended to a demand vector field theory for a population to represent population space-time travel patterns. We propose “vector kernel densities” as estimates of this field from observed activity trajectories. We show how these vector kernel densities can be used for time geography, particularly in 1) visualizing and monitoring the relationship between transportation systems and daily activity patterns that are streamed from a population, or 2) on the effect of changes to those systems on the patterns. As a result, we can inform policy-makers and the public on operational welfare effects (even in real-time, if the data was available) on a region based on events measured purely from Big Data and time-geographic urban informatics.

The study is organized as follows. Section 2 provides a literature review of vector fields in GIS, particularly on Miller and Bridwell’s (2009) methodology. Section 3 presents our proposed density-based methodology for representation population aggregations of vectors in space-time and applications in visualizing relationships with transport systems. Section 4 is an illustration of the methodology using survey data from the Greater Toronto Area. Section 5 concludes.

2. LITERATURE REVIEW

The review is broken down into 1) an overview of time geography and the need for vector-based visualization tools, and 2) a survey of recent developments in vector field-based GIS techniques.

2.1. Time geography and modelling of activity patterns

Time geography was introduced by Hägerstrand (1970) to measure individuals’ allocation of space and time (for time allocation, see Becker, 1965) throughout a day. Central to this theory is the conflict that arises from individuals seeking to maximize their utilities while constrained in time and space. Researchers have long realized the direct connection between time geography and travel behaviour, and have developed activity-based travel forecast models that align with the theory (e.g. Jones, 1979; Recker et al., 1986a,b; Ettema et al., 1993; Garling et al., 1994; Recker, 1995; Kitamura et al., 1996; Miller and Roorda, 2003; Arentze and Timmermans, 2004; Bhat et al., 2004; Chow and Recker, 2012; Chow, 2014).

A number of studies have considered GIS tools to visualize and measure activity patterns of travellers, e.g. Miller (1991), Golledge et al. (1994), Kwan (2000), Pandyala et al. (2002), Kwan and Lee (2003), Builiung and Kanaroglou (2006), Neutens et al. (2008), Miller and Bridwell (2009), Demšar and Virrantaus (2010), Chen et al. (2011), Goodchild (2013), Chen et al. (2013). Techniques of visualizing the behavioural aspects of travel have primarily fallen into two groups: the first is the use of activity prisms introduced by Hägerstrand (1970) and implemented in a GIS environment by Miller (1991). While this visual representation conveniently captures the constraint-based nature of time-space allocation decisions, Pandyala et al. (2002) demonstrated the challenges of estimating such a prism from travel data. These challenges include the lack of identification of the vertices (anchors) of the prism for an individual, and the uniqueness of the prism to each individual making it difficult to extrapolate to other individuals in a population. As a result, while activity prisms are helpful in understanding...
space-time trade-offs for an individual, it is less meaningful when trying to depict the patterns of a population and relate that to factors from the built environment.

A second approach focuses on population aggregations of travel patterns as density patterns in space and time, e.g. Kwan and Lee (2003), Chen et al. (2011), Demšar and Virrantaus (2010), Downs (2010), Goodchild (2013). In these studies, population-based static attributes or travel paths are used to form kernel functions to obtain densities for a population. In effect, these studies look at densities of trajectories using 3D kernel functions as opposed to 2D kernel functions. These studies nonetheless treat the shapes as static magnitudes without any directionality. A GIS visualization that includes both magnitude and direction can more effectively describe the patterns within the region. For example, a cross section of the path-based densities proposed by Chen et al. (2011) for a particular time would reveal a time-dependent density map of locations in space. If this visualization included direction, then the same cross-section would not only reveal locations in space but also the momentum of the densities.

Despite the number of studies on visualizing activity patterns, there are also very few time-space GIS methods to visualize the impacts of different transport system designs on the time-space patterns of a population. State of the art methods in transit system visualization and evaluation (e.g. O’Sullivan et al., 2000; Lei & Church, 2010; Langford et al., 2012) do not yet consider their explicit interaction with travellers’ time-space paths or prisms.

2.2. Vector field-based GIS
Miller and Bridwell (2009) propose a time geographic field theory as a more generalized form: conventional GIS assume isotropic cost fields that use scalar values that depend only on location, $k(x)$, while a vector cost field is anisotropic with a direction-specific cost function, $k(x, x')$. The term $x'$ is the velocity vector. This continuous, vector-based cost function is analogous to a flow or a current (e.g. electromagnetic fields). Miller and Bridwell (2009) introduced the anisotropic cost field case based on Puu and Beckmann’s (1999) continuous space concept, shown in Eq (1).

$$C_p(t_i, t_j) = \int_{t_i}^{t_j} k(x(t), x'(t)) \| x'(t) \| dt$$

(1)

where $\| x \|$ is a vector norm, and $C_p$ is the total cost of path $P$ from time $t_i$ to time $t_j$. In their study, they illustrate using velocity fields and the shortest paths through those fields to derive more nonlinear (and generalized) activity prisms for individuals. In other words, the prism can be constructed from minimum cost curves between points through an inverse velocity field, shown in Eq (2).

$$t^*_{kl} = \int_{P^*_kl} v^{-1}(x, x') dx$$

(2)

where $t^*_{kl}$ is the minimum travel time between two locations, $P^*_kl$ is the minimum cost path, and $v^{-1}(x, x') = \left[ \frac{dt}{dx_1}, \frac{dt}{dx_2}, ... \right]$ is the anisotropic inverse velocity field. The prisms are then identified in time geographic measurement theory as follows: the future field is $f_l(x_k) = t_l +$
the past field is \( p_j(x_k) = t_j - t^*_{kj} \), and the potential path field is \( g_{ij}(x_k) = (t_j - t_i) - (t^*_{ik} + t^*_{kj}) - \alpha(x_k) \), where \( i \) is associated with anchor \((x_i, t_i)\) and \( j \) with anchor \((x_j, t_j)\) (Miller and Bridwell, 2009).

There are not many vector field techniques proposed in the GIS literature, particularly within time geography. Zhong et al. (2012) showed vector fields for water flow (and only in 2D space) and suggested using colours to identify direction of vectors, but they did not consider applications in time geography. A few studies have considered the velocities of moving entities (Noyon et al., 2007; Orellana & Wachowicz, 2011), but not with the formal theoretical treatment given by Miller and Bridwell (2009).

3. METHODOLOGY

While Miller and Bridwell (2009) used anisotropic inverse velocity fields to construct activity prisms for individuals, we propose using vector fields in a different manner and for a different purpose. We propose demand vector fields (as opposed to inverse velocity vector fields), which are direction-dependent vectors extracted from observed travel trajectories.

3.1. Demand vector fields

Consider a 3D continuous space domain in which a point is characterized by the tuple \((x, t)\), with \( x \in \mathbb{R}^2 \) and \( t \in \mathbb{R}^+ \). By default, \( t = 0 \) is assumed to be the start of a day. The space is populated by a set \( N \) of individual trajectories.

**Definition 1.** An individual trajectory, \( P_n \), is a monotonically increasing path by individual \( n \in N \), from \( t_i \) to \( t_j \).

**Definition 2.** A local demand vector field is an everywhere differentiable and nonnegative anisotropic function that assigns a numeric measure of local demand along a direction for an individual \( n \in N \), \( k_n(x(t), x'(t)) \), where \( x'(t) > 0 \) is the derivative of the path function, and is always positive in time (no freezing in time or going backwards in time). A vector \( k \perp x \in \mathbb{R}^2 \) is considered demand for activity participation. A vector \( k: k \cdot x > 0 \) is considered demand to travel.

**Definition 3.** The total travel demand of individual \( n \) for path \( P_n \), \( d_n(P_n) \), is defined as the path integral in Eq (3).

\[
d_n(P_n) = \int_{t_i}^{t_j} k_n(x(t), x'(t)) \|x'(t)\| dt
\]

**Definition 4.** The local population demand vector field is an everywhere differentiable and nonnegative anisotropic function that assigns a numeric measure of local demand along a direction for a population, \( K(x(t), x'(t)) \) and shares similar properties as the individual local demand vector field.
**Proposition 1.** The local population demand vector field is shown in Eq (4) and monotonically increases in time.

\[ K(x(t), x'(t)) = \sum_{n \in N} k_n(x(t), x'(t)) \]  

(4)

**Proof.** The population density field \( K(x(t), x'(t)) \) is a vector sum of the individual vector fields \( k_n(x(t), x'(t)) \), which results in Eq (4) by definition. The sum is non-negative in time because the individual vectors are monotonically increasing in time. Vector summation is illustrated in Figure 1.

![Figure 1. Illustration of vector summation to obtain population demand vector field.](image)

**3.2. Interpretation and applications**

The local population demand vector density field can be interpreted as the aggregate demand of a whole population to travel at a specific point in the space-time domain.

**Definition 5.** The total population travel demand for path \( P \), \( D(P) \), is defined in Eq (5).

\[ D(P) = \int_{t_i}^{t_f} \sum_{n \in N} k_n(x(t), x'(t)) \|x'(t)\| dt \]  

(5)

Since individual vectors are assumed independent of other vectors, the population vectors can be segmented by demographics or other classifiers.

The population travel demand for a particular path is useful because different transport systems can be modeled as paths or surfaces in space-time, as shown in Figure 2. The first one shown in Figure 2a is road infrastructure that travelers can access at any time of the day. Because there is no time constraint, the system is represented by a surface in time, \( S \), which becomes a curve when projected onto \( x \). Figure 2b illustrates fixed schedule transit lines that run repeatedly in time. A person trying to board this type of system can only do so at specific times...
and locations. As such, these scheduled service systems are represented by paths in time, which become curves when projected onto $x$ for fixed routes.

**Figure 2.** Illustration of (a) road infrastructure as a surface in time, and (b) a fixed schedule transit line with a constant headway.

**Definition 6.** A transport system is a surface $S$ comprised of a vector field $k_S(y(x, t), y'(x, t))$ such that projection on $x$ results in a static infrastructure curve $S_0$, where $S_0 = (S, x)$. A transport system with scheduled access is a monotonically time-increasing path and is denoted by a time index, $S_t$. The path derivative $y'(t)$ is the velocity of the transport system.

**Proposition 2.** The population demand for a transport system, $D_S$, is the positive projection of the population demand vector field onto the transport system surface at the same location, $D_S = \langle K(x, x'), k_S(y(x), y'(x)) \rangle_+.$

**Proof.** Population demand for travel that goes in a negative direction from that of a transport system’s path vector would lead to a negative demand for the system. Since negative demand cannot exist, only positive projections are considered. An illustration of the projections is shown in Figure 3.

Note that the $D_S$ is not a forecast of demand for a transport system, since it is not fitted to actual ridership numbers. Instead, it is regarded as a visual correlation between two sets of measures. It may be possible to identify causality and develop some type of forecast model based on $D_S$, but that is beyond the scope of this study.

In some cases, the demand for travel is at a higher speed than that offered by the transport system. This difference, whether it’s positive or negative, might be a useful attribute for identifying correlations with wait times and departure time choices, as well as making predictions of mode choices based on alternative system lines and the demand vector field. This will be examined more carefully in future research.
Another application of the population demand vector field is to compare the fields of two different scenarios or time periods, such as for a before-after study. For example, suppose there was a major storm or other event that led to a significant shift in travel and activity patterns on a certain day. If we know the population demand vector field on the day prior to the event as well as for the day of the event, the impacts in terms of travel demand can be measured directly.

**Definition 7.** The *field differential* between two demand vector fields $K_1$ and $K_2$ is itself a vector field that can be positive or negative, and is determined by Eq (6).

$$\Delta(K_1, K_2) = K_2 - K_1$$

Since the fields are vectors, subtraction is simply a vector operation. The resulting vector field shows the regions in space-time where major changes result. This technique can measure the effects of events on a population as well as longitudinal trends such as population growth, peak hour spreading due to congestion, or the growth of new regions.

### 3.3. Kernel density estimation of population vector field from observed trajectories

In practice under a Big Data setting, $P_n$ are observed from data sources like travel survey diaries, GPS data, or other data sources and sensors. In these cases, the estimated local demand vector field is denoted by $\hat{k}_n$. If only a cross-sectional data set is obtained, there may only be one observed path per individual. As a result, the estimated individual local demand vector fields do not provide a complete picture for the local demand vector field for the entire domain of that individual.

In this situation, we estimate kernel densities of the population vectors based on the estimated $\hat{k}_n$ from observed trajectories. For convenience, the kernel density estimation method from Silverman (1986) used in ArcGIS is adopted.
**Definition 8.** For a given sample set $\Lambda$ of trajectories $P_{\text{MEA}}$, let the estimated vector kernel density $\hat{f}(x(t), x'(t)|P_{\text{MEA}})$ be an estimation of the population demand vector field. Unlike a kernel density of a scalar variable, the vector kernel density $\hat{f}$ captures both the magnitude and the direction of the demand.

### 3.4. GIS implementation

To generate a vector kernel density as an estimate of the population demand vector field, we apply three steps.

**Step 1 – discretization of domain $(x, t)$**

Firstly, we need to decompose continuous large-scale 3D time-geographic space into quantitative units. Suppose there is an artificial square city with a spatial resolution of 50 by 50 meters. We can create 2D grids to represent the continuous geographic space in the artificial city, as shown on the left in Figure 5. If we add time dimensional resolution (such as one minute) onto the 2D geographic representations, then we can get the 3D cubes to represent the continuous 3D geographic space (on the right in Figure 5). This is the fundamental way that we decompose the continuous and large-scale geographic space to analyze our study area using the vector- and raster-based geographic data in a quantitative manner.

![Figure 5. 3D cubes as geographic representation.](image)

**Step 2 – extraction of observed trajectories to estimate $\hat{k}_n$**

Based on the above geographic representations, we intersect trip vectors with 3D cubes in Figure 5. Firstly, a trip vector (in blue on the right of Figure 6) is split into several ones. In Figure 6 on the left, the trip vector is split into $V_{(t_1, t_2)}$, $V_{(t_2, t_3)}$ and $V_{(t_3, t_4)}$, where each of the split vectors is valid for a specific time layer of 3D cubes. For example, $V_{(t_1, t_2)}$ is valid for the set of 3D cubes starting from time $t_1$, and if we intersect $V_{(t_1, t_2)}$ with these 3D cubes, only $C_{11}$, $C_{12}$, $C_{13}$, $C_{14}$ and $C_{15}$ will be affected. As there are many trip vectors that intersect with the cubes in different time layers, a single 3D cube can be affected by one or more split vectors. We add up split vectors in each cube using vector addition (up-right in Figure 6) to count the total contribution of all trip vectors.
**Step 3 – estimate kernel density function $\hat{f}(x(t), x'(t) | P_{nE})$**

The last step is to create a density map for each time layer and vectorize each cell. The way to create a scalar density map is straightforward, such as Kernel Density Estimation in 2D space using ArcGIS, because cubes in each time layer have values according to the aggregated vector in each. With regards to assigning directionality to the cubes in the generated density map for the whole geographic space, we use a given radius for each aggregated vector to determine the number of cells that are affected. The direction of these cells within the given radius point to the end of the aggregated vector, and the influence of the aggregated vector on such cells decrease along the radius. **Figure 7** shows the process. On the right of **Figure 7** is the result of cell vectorization from one aggregated vector, with the red showing the most affected and the blue showing the least affected. Similarly, we can add up the directions in a cell affected by different neighboring cells.

![Figure 7. Vectorization of density map.](image-url)
4. CASE STUDY: GREATER TORONTO AREA

A case study is conducted of the Greater Toronto Area (GTA) to demonstrate the effectiveness of the proposed methodology and applications.

4.1. Data description

The methodology is tested using real travel survey data obtained from the Data Management Group (DMG, 2011) at University of Toronto. The Transportation Tomorrow Survey is conducted every five years; we obtained the survey data for 2011 and 2006 for this case study. The data is based on user diaries with zone-masked (Figure 8a) location information in the GTA. It includes personal, trip (Figure 8b), transit, and household information (DMG, 2011).

![Figure 8](image)

(a) 2,272 zones in Great Toronto Area in red points, and (b) 624,845 trips of 311,022 persons from 118,280 households in the year 2011.

Interestingly, we find that the number of activities in the zones demonstrate an imbalanced distribution (on the left in Figure 9). The mean value of number of activities in a zone is 550. We group persons who have activity zone IDs together, and there are 57,051 of them. The group sizes also exhibit an extreme imbalance as shown in Figure 9 (right). The top group sizes are 7692, 1448, 1325 …, and the smallest size is 1. There are only 3,578 (5%) groups whose sizes are greater than 3 (the average size of all groups). The imbalanced distribution can also be found in the trip lengths, and all of them follow a kind of heavy-tailed distribution. This universal phenomenon captures the feature of complex urban systems, and it makes sense since a transit system is supposed to be a sub-system of an urban system. Assuming that the samples in our dataset represent the whole population in our study area, the imbalance of how people use the urban space will impose an inevitable influence on how we evaluate and understand the relationships between activity patterns and transport systems.
4.2. Vector kernel density maps

By applying our methods on these data sets, we create 50 by 50 meter spatial grids in the GTA study area, resulting in over 3.5 million of them. To capture the most detailed activity patterns, we select one minute as the time resolution to create 3D cubes to represent the 3D space in GTA. There are 24 hours per day, and therefore there are 1,440 minutes (or time layers) per day. In total, there are 1,440 * 3,500,000 = 5,040,000,000 cubes. Although it is not always necessarily to create all the cubes physically, current GIS software either could not handle such size or becomes extremely slow when we process such data using our methods. Instead, we create a simpler geometric data model and build up both the 2D and 3D index to speed up the computation throughout our work.

We intersect the 624,845 trips with all 3D cubes through 1,440 time layers and assign trip vectors to cubes as described in Figure 6. For all the cubes in each time layer, some of them intersect with one or more trip vectors, while others do not intersect with any other trip vectors. After assigning values of added trip vectors in the cubes, we create non-vertical (activity participation segments are left out) vectors representing travel density. Figure 10 shows the vector kernel density of one time at 8:00 AM in 3D scene.

Figure 9. Rank-size distribution of activity number in zones (left) and group sizes (right)

![Figure 9. Rank-size distribution of activity number in zones (left) and group sizes (right)](image)

Figure 10. Vector kernel density at 8:00 AM in the year 2011 in Great Toronto Area, where the red shows the highest density and green shows the lowest density in 3D space.

![Figure 10. Vector kernel density at 8:00 AM in the year 2011 in Great Toronto Area, where the red shows the highest density and green shows the lowest density in 3D space.](image)
The generated vector kernel density map is in raster format. For each cell in the raster density map, we use the added or aggregated vectors in the cubes to vectorize all the cells according to the method described in Figure 7. Each cell in the density map has both a scalar density value and a vector indicating the direction. If we visualize the vectors in raster cell, we get the vector density map as shown in Figure 11 in 2D space. From the legend we can see that the colour and arrow means the direction, while the width of the arrow indicates the density value: the wider the arrow, the higher the density.

We can combine the two visualizations (i.e., Figure 10 and Figure 11) together into one to make a vector density map in 3D space as show in Figure 12, where the colour shows the direction while the height is the magnitude of the density. On the left is the view of GTA from the south (Lake Ontario) and on the right is the view from the north direction. From the density height we can see that there are some peaks indicating the downtown Toronto and some smaller sub-city downtown areas such as Oshawa and Hamilton, while from the direction colours we can see that most of the people are moving to these downtown areas at 8:00 AM in the morning.
4.3. Projection of demand vector fields onto a sample transit service line

We project the vector densities onto polylines belonging to specific transport systems (like a bus line) to visualize the relationship between activity densities with a particular transportation system.

We also identify peak/off-peak thresholds for an object; we divide the projected data over time (instead of summing over all time) to see how the projected density varies by time of day. As noted earlier, such a projection is not meant to forecast realized demand for a system; it is instead a visual measure from which forecast models can be based. We select the eastbound bus #506 (Figure 13) operated by Toronto Transit Commission (TTC) in Toronto, partly because it goes through the downtown area with a high possibility to capture interesting patterns.
Figure 13. The transit line and stops of eastbound TTC bus #506 in Toronto.

The blue line in Figure 14 is the density along one loop line of eastbound #506 that starts from 8:00 AM in 2011. Further, we can demonstrate the effect that relocating stop locations can have on the projected demand densities. For this purpose, we have a second scenario where we relocated the original stops to equal distances apart on the line. Accordingly, we can get the density of relocated stops (orange line in Figure 14). We conjecture that the high slopes along the profile can indicate potential sources of bottlenecks.

Figure 14. Vector density projected on the transit line of Bus #506: the blue line is the density of original stops and the orange line is density of the equidistantly relocated stops.
This methodology provides a means of measuring the accessibility (from both a space and schedule perspective) of a population to a transport system. The coverage can be measured by the vector densities within a polyhedral space formed by the transport system over time. The coverage or accessibility area can be calculated by using the following definition:

\[
Coverage = \sum_{k=0}^{n} W_k D_k
\]  

Where \( n \) is number of stops minus one, \( W_k \) is the width between two consecutive stops and \( D_k \) is average density of such two stops. This process can be illustrated in Figure 15 by taking the original stops as example. Suppose the coverage of original stops is \( C_o \), and the coverage of equidistantly relocated stops is \( C_r \), then the rate of accessibility difference between two set of stops can be evaluated using the following definition:

\[
Rate = \frac{C_o - C_r}{C_r}
\]

![Figure 15. Calculating the areas of density for transit line as the indicator of accessibility.](image)

4.4. Comparison between 2006 and 2011 using vector field differential

We showcase the methodology as a means of computing the differentials in activity densities for before-after studies of changes to a transport system (e.g. investments or disasters). We accomplish this with densities estimated from 2006 and 2011 Transportation Tomorrow Survey...
results (DMG, 2011). This illustration provides guidelines for real-time measurements of changes in activity utility and travel disutility due to shifts from one state to another, with implications for visualizing and quantifying disaster effects on a population’s economic welfare as measured directly by activity behavioural patterns. Figure 16 shows time slices taken from the two years.

Figure 16. Vector density map at 8:00 AM in the year 2011 in Great Toronto Area: views from south (on the left) and from the north (on the right).

For the field differential, Figure 17 (top) shows the result where the positive parts indicate the increased local population demand from 2006 to 2011, while the negative parts imply a decreased local population demand. Figure 17 (bottom) shows the differences in direction (red is positive along the direction of the arrow, while blue is negative along the direction of the arrow) between the year 2006 and 2011.
Figure 17. The difference of densities (top) and vector densities (bottom) at 8:00 AM between the year 2006 and 2011 in GTA. (Note: the blue arrow means the negative differences, while the red arrow means the positive differences).
5. CONCLUSION

The rise of Big Data has made it possible to use demand data at an operational level, which is necessary to directly measure the economic welfare of operational strategies and events. In order for that to happen, we argue the need for more adequate visualization and quantification method. GIS is the primary visualization tool in this regard, but most current methods are based on scalar objects that lack directionality and rate of change—key attributes of travel. The few studies that do consider vector-based GIS have largely looked at vector fields for individuals, not populations.

In this study, we extend upon that literature by introducing vector field for a population: the population demand vector field. We show how to estimate this field using a vector kernel density estimated from observed trajectories of a sample population. By representing transport systems as vector fields that share the same time-space domain, we can use projections to visualize relationships between population demand and these systems. This visualization tool offers a powerful approach to visually correlate changes in the systems with changes in demand, as demonstrated in our GTA case study. The significance is that, in the age of Big Data, we can potentially measure in real time the effects of disasters on the economic welfare of a population, or quantify the effects of operational strategies and designs on the behavioural activity patterns of the population.

There are a number of studies that can be conducted in the future. We believe we have only scratched the surface in the applicability of this methodology to visualize system changes on real time demand patterns. While the TTS data that we used was adequate to illustrate the methodology, its real capability to monitor changes in real time can only be witnessed in an operational setting by implementing it in GIS package designed for a video wall platform with real time feeds. This opens the door for transportation operations beyond traffic control as control centres can also be used to monitor real time demand (or planning agencies can create departments for operational planning that monitor real time demand). Empirical studies based on actual events and measurement using the methodology would give credible evidence of its use. Short-term forecast models based on the vector fields from this methodology may offer researchers new insights.

ACKNOWLEDGEMENTS

This research was undertaken, in part, thanks to funding from the Canada Research Chairs program and an NSERC Discovery Grant. The authors wish to thank the University of Toronto’s Data Management Group for the Transportation Tomorrow Survey data used in this study.

REFERENCES


DMG, 2011. Transportation Tomorrow Survey, Data Management Group, University of Toronto.


Understanding the Usage of NYC Taxicabs for Vehicle Electrification in a Large Metropolitan Area

Promiti Dutta and Roger N. Anderson
Center for Computational Learning Systems
Columbia University
New York, NY
Abstract
The term “big data” is a generalized phrase applied to any large data set that is difficult to process using traditional data processing applications. Challenges around big data include capture, search, retrieval, sharing, analysis and visualization. In this paper, we introduce a novel data set captured by the New York Taxi and Limousine Commission in 2012 to describe the usage of taxicabs in a large metropolitan area, New York City. This unique data set is important in urban informatics for adequate transportation planning in metropolitan areas. Issues such as the number of taxis needed on the roadways to the feasibility of converting taxicabs into electric vehicles for environmental sustainability can all be investigated through the acquisition of such data. As a result, this data set is useful for public policy planning (i.e. how many medallions should be available per year in a metropolitan area) to economics related concerns of revenues available from taxicab usage in cities to road and infrastructure planning by civil engineers and urban designers. In this paper, we describe one potential application and usage of this data through the creation of an agent based model for simulating vehicle-to-vehicle (V2V) wireless charging for vehicle electrification in large metropolitan areas.

Keywords: big data, electric vehicle, taxicabs, vehicle electrification

Introduction
Urbanization is the process of increasing the number of people that live in urban areas. In 2012, the population of New York City (NYC) was approximately 8.3 million people, making it the most populous city in the United States and amongst the most populous urban areas in the world. New York City accounts for approximately 2% of the total population of the United States, with Manhattan accounting for 19% of the total New York City population. An urban setting with such great populations also has high transportation needs.

Vehicles, both personal and commercial, have become a ubiquitous form of transportation in the developed world. While, public transportation (subways) is the mass transit method for transportation for many, NYC still has many cars on its roadways adding to the air pollution levels. Many of these cars are additional transportation means for the population including buses, shuttles, para-transit, and taxicabs to list a few.
In 2011, President Obama announced in his State of the Union address that his administration would push to have 1-million electric vehicles on the road by 2015. The auto industry is amidst a technological transformation in identifying alternative sources of energy to power vehicles due to two driving forces: environmental pollution prevention and depletion of fuel resources. This drive for developing "smarter" solutions to create a "smarter planet" is crucial to advancing the technological science of electric vehicles (EVs). As alternative methods for creating energy are being sought, we see an increased interest in electric vehicles as one potential solution for lessening our dependence on fossil fuels. Currently, the main hurdle with adoption of EVs is due to their limited driving range, lacking infrastructure for recharging and time needed to recharge.

Among potential early adopters for EVs, are fleet-based vehicles. Fleet vehicles are defined as groups of vehicles that are owned or leased by a business as opposed to an individual or family. Types of fleet vehicles range from cars to vans to trucks, depending on the need of the company. Multiple drivers, multiple paths, or any combination of the two can use any single vehicle in a fleet system. In essence, fleet vehicles can be thought of as moving micro-grids or networks with nodes and parameters. Fleets of vehicles can be clustered into groups: structured and unstructured fleet vehicles (Table 1). Each of the types of fleet vehicles serves very different functions.

In this paper, we focus on the implications of converting taxicabs in a large metropolitan setting into EVs. We quantify the current usage and driving behaviors of taxicabs by using a unique big data set from the New York City Taxi and Limousine Commission. The paper is outlined to discuss the availability of the data set, a detailed description of the data set, data analysis conducted to understand driving patterns, and finally application of the data set to determine the feasibility of vehicle electrification for an all electric taxicab fleet.

<table>
<thead>
<tr>
<th></th>
<th>Structured</th>
<th>Unstructured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schedule</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Network path</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Example</td>
<td>FedEx, UPS, MTA</td>
<td>Taxicabs, Rental Cars, Commuter Vehicles</td>
</tr>
</tbody>
</table>
Data Set Availability

The New York City Department of Taxis and Limousines (NYC TLC) captures data about its medallion owned taxicabs. This data serves multiple purposes for the organization and is available to the public via Article 6 of the New York State Public Officer’s Freedom of Information Law (FOIL). The Department of Transportation (DOT) does not make the data readily available via any mass dissemination methods (i.e. websites), but allows people to acquire access to copies by bringing brand-new unopened hard drives to their offices. The files are available as comma separated value (csv) files split by month with each file containing approximately 2.5 gigabytes of data for the trip data and 1.75 gigabytes of data for the corresponding supplementary fare data. The total data set inclusive of both trip and fare data is approximately 50 gigabytes.

As part of New York City’s initiative to make issues more transparent and efficient for the public viewing, there has been a bill introduced entitled, the OpenFOIL bill, which will create a more transparent and responsive online system to make information and data more readily available. Many FOIL requests are ignored or remain not responded to due to lack of adequate resources. Additionally, OpenFOIL contains an extremely important data provision that states that once a data set is has been made available to one person, it will be included in the city’s open data portal. The availability of a system to dissemination public information will be extremely useful for urban planning and research methods. Currently, this centralized dissemination of data is common practice elsewhere in cities such as Oakland and Chicago.

Data Set Description

The data set contains over 192 million data entries for each paid ride in a medallion owned taxicab from January 1, 2012 to December 31, 2012. The relevant information contained in the primary data set is summarized in Table 2. A supplementary data set that links to this original data contains payment information for the taxicabs. It is de-identified by medallion number and drivers license ID, but keeps track of the fare paid and the type of payment used to pay the fare.
<table>
<thead>
<tr>
<th>Data Field</th>
<th>Data Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medallion number</td>
<td>String</td>
<td>De-identified medallion number for taxicab.</td>
</tr>
<tr>
<td>Drivers license ID</td>
<td>String</td>
<td>De-identified drivers license ID since a medallion can have multiple drivers.</td>
</tr>
<tr>
<td>Pick up date and time</td>
<td>Date and time</td>
<td>Describes the date and time at which a passenger’s meter started measuring the fare. This does not necessarily the exact moment at which the passenger entered the vehicle as cab drivers often press the button after the cab has already started driving.</td>
</tr>
<tr>
<td>Drop off date and time</td>
<td>Date and time</td>
<td>Date and time at which a passenger’s meter has stopped measuring the fare. Does not indicate the exact moment the passenger has reached their destination as cab drivers can deactivate the meter before arriving at the destination to offer the passenger time to provide payment earlier for quicker drop off.</td>
</tr>
<tr>
<td>Passenger count</td>
<td>Number</td>
<td>Describes the number of passengers that were in the vehicle. This number is often inaccurate since fares in NYC are not dependent on the number of passengers. As a result, many cab drivers do not modify this number during a passenger trip.</td>
</tr>
<tr>
<td>Trip time in seconds</td>
<td>Number</td>
<td>Calculated trip duration from pickup date and time to drop off date and time. For the reasons mentioned above, this trip duration number is not fully indicative of the total trip duration.</td>
</tr>
<tr>
<td>Trip distance in miles</td>
<td>Number</td>
<td>Calculated based on the number of miles registered by the taxicab from the time pickup was initiated to drop off time. This number may be slightly skewed due to the reasons mentioned above regarding accuracy of measurement.</td>
</tr>
<tr>
<td>Pick up longitude</td>
<td>Number</td>
<td>Longitude coordinates where the passenger was picked up. More specifically coordinates at which point the meter was turned on.</td>
</tr>
<tr>
<td>Pick up latitude</td>
<td>Number</td>
<td>Latitude coordinates where the passenger was picked up. More specifically coordinates at which point the meter was turned on.</td>
</tr>
<tr>
<td>Drop off longitude</td>
<td>Number</td>
<td>Longitude coordinates where the passenger was dropped off. More specifically coordinates at which point the meter was turned off.</td>
</tr>
<tr>
<td>Drop off latitude</td>
<td>Number</td>
<td>Latitude coordinates where the passenger was dropped off. More specifically coordinates at which point the meter was turned off.</td>
</tr>
</tbody>
</table>
The data also contains information about the amount of tolls and tips paid to the driver. This information is seemingly accurate for credit card purchases as the system registers the difference between the fare and total amount paid by the passenger, however, in cases of cash payments, there is reporting bias present in the data set since taxicab drivers oftentimes do not key in the total cash amount tendered into the system. Additionally, if a cash tip was made on top of a credit card payment, the system does not have a method to record this information.

While the data provides a detailed and in-depth glimpse of how taxicabs are utilized in the city, the data quality is not completely accurate. Much of the information available in the data set is retrieved automatically based on the turning on and off of the meter during a paid passenger ride adding an inherent reporting bias to the data. The time at which a cab driver turns on the meter varies for each cab driver. For example, certain cab drivers, will turn on the taxi meter prior to stepping out of their vehicles to help load bags into the trunk, while others will start driving the vehicle instantaneously after the passenger is seated and will turn on the meter as the drive is ongoing. In certain instances, the meter takes a while to reset from the last ride or has to turn on when a shift has just begun for the taxicab driver and causes a longer delay as to when the data is acquired.

**Data Analysis**

There are several types of analysis that can be performed on a large data set. The NYC TLC FOIL data has endless possibilities in characterizing usage in a large metropolitan area ranging from simple analysis of the time and travel distance of a typical taxi ride to more involved calculations that seek the hourly/daily/monthly usage pattern of taxicabs. Analysis has also been done by the TLC to understand the usage patterns with relation to the origin and destination of paid rides. To that end, more than 95% of taxi trips begin in Manhattan or at the local airports.

Taxis in New York City operate in shifts with two daily shift changes occurring in the early morning hours and late afternoon hours. Coincidentally, there is a shift change on during the typical weekday evening commute. However, there are negligible differences in the driving distances between taxicabs during morning and afternoon shifts.
We determine the number of taxicabs active on the streets for all hours during the day, with 0 indicating midnight through 23 representing 11 PM (Figure 1). According to the figure there are varying margins of availability based on day year, but certain times regardless of day of week or month there are very small margins such as rush hour traffic around 4PM. From the TLC, we understand that most shift changes occur between 4PM and 5PM and again between 4AM and 5AM daily. Shift changes require the driver to return the taxicab to its assigned garage. Failure to arrive by a certain time can result in a fine for the taxicab driver. In fact, the data shows that the typical number of taxicabs available on the streets of New York between 4PM and 5PM diminished by approximately 20% each day. Coincidentally, this shift change occurs during one of the highest demand times for taxicabs for the evening commute.

There is also a gradual decline in the number of taxicabs available each day between midnight and 5AM. While this corresponds to the typical shift change problem between 4AM and 5AM, the demand for taxicabs during these early hours is also limited. Hence, many taxicabs end their shifts early. The usage of taxicabs during these hours are increased during certain holidays (i.e. New Year Eve and Day) as well as weekend when a nightlife is more prominent in the city.

From preliminary analysis of the data set, we note an interesting pattern in taxicab usage throughout the year, in which the largest driving distances for taxicabs is at the beginning of the year from January – February and then tapers to a constant range for the rest of the year (Figure 2). This could be due to the colder temperatures during that time in the city coupled with increased visitors in the area.

However, analysis of taxicab usage by number of paid taxicab rides as opposed to distance traveled shows that the in the 2012 year, there were a 178,544,324 rides with an average of 14,878,693 rides per month and standard deviation of 638,965 taxicab rides. Figure 3 summarizes the number of rides per month. In this case, we see that March has the most number of paid taxicab rides and November has the least.
Figure 1: Number of active taxicabs on the streets of New York based on hour of day for the 365 days in 2012. Each color represents a different day.

Figure 2: Scatterplot showing variation of paid driving distances by date. Notably, the beginning of the year has larger cumulative driving distances than the rest of the year. The larger distances are likely due to increased taxicab rides to the area airports and inclement weather usage.
Applications of Data

There are several applications for this data that range from short-term service solutions to long-term fleet conversion issues. Service solutions include better methods for reducing the time and distance between taxicab pickups for passengers thereby addressing two issues: (1) allowing the taxicab driver to have more paid passengers during a shift and (2) having better coverage and service for taxicab passengers to reduce wait time in finding a taxicab. This could be accomplished by learning and creating models from the data about drop off and pick up locations to better provide both passengers and taxicab drivers better ideas of where to find taxicabs during different times of the day. Fleet conversion issues center around the desire to convert taxicabs into electric vehicles. In this section, we focus on the challenges around this conversion.

Determining Feasibility of All-Electric Taxi Fleet

Driving distance varies for a vehicle based on the type of vehicle. Commuters use their vehicles very differently from service fleet vehicles. Using the data, we extrapolate the minimum and maximum cumulative distance a taxicab may have driven during a 12-hour shift. Cumulative distance is the total distance that a taxicab drives both with a paid passenger as well as while in

Figure 3: Total number of paid taxicab rides per month in 2012. March has the most number of rides, while November has the least. The average number of rides per month is more than 14 million.
search of the next passenger. We estimate a minimum cumulative distance and a maximum cumulative distance.

For minimum distance, we use the GPS coordinates between the drop off and pick up points to calculate a “Manhattan” street distance between the two points by assuming that the car drives straight to reach the desired latitude and then turns and drives straight to reach the desired longitude. This calculation is slightly skewed as all streets in Manhattan do not run in both directions and this method of driving may not be the most efficient in the crowded streets.

To calculate the maximum distance, we compute the time elapsed between the drop off and pick up of passengers. We extrapolate a distance driven during this time with an assumption that the taxicab traveled at the 30 miles per hour speed limit during this time interval. Figures 4 and 5 show the minimum and maximum cumulative distance driven by a taxi that can be fitted to a heavy-tailed Rician distribution.

Given these driving distributions for taxicabs in New York City and the known possible ranges of driving distances in common EVs (approximately 75 miles according to Environmental Protection Agency), only 18.8% - 20.7% of NYC taxis would be able to complete an entire taxicab shift without needing to recharge. If we increase the driving range of electric vehicles to 150 miles, approximately 50.7% - 69.4% of taxis would be able to complete a taxi shift. However, if the driving range of electric vehicles increases to 300 miles, then 99% of all taxis would be able to complete a 12-hour shift without needing to refuel. Of course the shorter the driving range of EVs, the less money that a taxicab driver can make during a standard 12 hour shift. To ensure desirability of EVs in the taxicab fleet, there is a need to ascertain longer driving distances for EVs without the need to frequently recharge. Current quick charging rates for EVs currently require approximately 30 minutes to regain 80% of the battery life, thus taxicabs may not be willing to have so much idle time.
Agent-based Modeling for Vehicle-to-Vehicle Wireless Charging

Using this data, we propose a novel vehicle-to-vehicle (V2V) wireless charging schema for vehicle electrification of electric vehicles (EVs). This ad hoc network, unlike current proposals for extending driving ranges of EVs, relies on EVs creating a social network of vehicles to assist each other with charging by essentially creating a moving network of charge stations. Since any vehicle can participate in exchanging charge with another vehicle either at traffic lights or by selecting a predetermined rendezvous point, we apply techniques from computer networks and communications to determine an effective method for EV routing and scheduling of charge transfers as well as game theory principles to determine incentives for participation through pricing. The availability of such a robust big data set affords us the opportunity to effectively model the usage of taxicabs in the city to determine the efficacy of a proposed system of V2V charge sharing.

We demonstrate the feasibility of the V2V schema via an agent-based model in which a moving network of charge sharing is created since each EV can act as a charge station. This “social network” of charge that is formed allow us to let EVs exchange charge with one another using IPT without needing to change any current infrastructure on roadway. We model a discrete-time simulation in which each iteration represents one unit time. During a unit time, an EV travels one
unit distance and expends 1 unit charge. We model a grid space to simulate NYC’s blocks, 250 units by 20 units. However, unlike streets in NYC, our streets assume two-way traffic.

EVs can communicate with one another such that they can identify and coordinate rendezvous points. When an EV enters the system, it has the following properties:

- Type of car: taxicab or commuter
- Charge, c: Cars enter the system with 100 charge units to represent a full tank.
- Distance to destination, d: The distance a car drives to reach its destination is randomly assigned using the real-world data.
- Start point
- End point

The ability to participate in charge exchange is determined by a simple heuristic: EVs only exchange charge when they have enough to reach their destination. The greedy approach helps to guarantee that cars that would have been able to reach their destination without recharging are still able to do so and only exchange their excess charge. Cars are willing to share all their excess charge.

We apply fisheye routing to coordinate rendezvous points between vehicles and simulate a 24-hour time period using 1440 iterations (24 hours/day * 60 minutes/hour = 1440 minutes = 1440 iterations). The general model for simulation involves having the cars enter the system and being randomly placed at a starting point (node i) with a destination also randomly assigned (node j). We apply Dijkstra’s algorithm to determine the shortest path between node i to j. Depending on whether a car has excess charge available or can reach its destination, we use the Fisheye Routing to reroute vehicles and coordinate rendezvous points.

To evaluate the feasibility of our system using the probability of refueling, we vary the ratio of taxicabs to commuter vehicles to further determine the tipping point at which an EV taxi fleet is sustainable. Figure 6 shows the probability of refueling for taxicabs during a 12-hour shift for various ratios of taxicabs to commuter vehicles. We consider refueling a failure as this would indicate that the taxi was not able to stay on the roads to serve its customer and is costing the taxicab driver time recharging and potential revenue. We also notice an exponential trend in the probability of refueling based on the ratio of taxicabs to commuter vehicles.
Figure 6: The probability of refueling decreases as the ratio of taxis to commuter vehicles increases. Implicitly this indicates that as the amount of excess charge increases in the system, the number of failures decreases.

Conclusions

The individual data set for 2012 is helpful in providing an insightful view into taxicab usage in a large metropolitan area. While there are many immediate applications for a data set like this, focusing on long-term modifications to the taxicab transportation system is benefitted with such data points. Using this big data set we are able to show a proof of principle model for urbanization of a large metropolitan area through conversion of taxicabs into EVs.

The power of a data set increases as the sample size increases. Longitudinal data will be very helpful in avoiding biases caused by unreliable data (i.e. determining whether the anomalies noticed in the 2012 data set are indeed anomalies or whether there is a pattern to these occurrences). While the 2012 FOIL data has many data points for the year, longitudinal data will help provide more data points for each day and collective month to better understand usage over time for taxicabs in a metropolitan area. Correlations to available weather data sets could also provide an unique understanding of different stimuli that affect the usage of this particular transportation method.
The Impact of Land-Use Variables on Free-floating Carsharing Vehicle Rental Choice and Parking Duration

Mubassira Khan  
The University of Texas at Austin  
Department of Civil, Architectural and Environmental Engineering  
301 E. Dean Keeton St. Stop C1761, Austin TX 78712-1172  
Tel: 512-471-4379, Fax: 512-475-8744  
Email: mubassira@utexas.edu

Randy Machemehl  
The University of Texas at Austin  
Department of Civil, Architectural and Environmental Engineering  
301 E. Dean Keeton St. Stop C1761, Austin TX 78712-1172  
Tel: 512-471-4541, Fax: 512-475-8744  
Email: rbm@mail.utexas.edu

Zhanmin Zhang  
The University of Texas at Austin  
Department of Civil, Architectural and Environmental Engineering  
301 E. Dean Keeton St. Stop C1761, Austin TX 78712-1172  
Tel: 512-471-4534, Fax: 512-475-8744  
Email: z.zhang@mail.utexas.edu

Word Count:

<table>
<thead>
<tr>
<th>Type</th>
<th>Count</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td></td>
<td>5822</td>
</tr>
<tr>
<td>Table</td>
<td>4 x 250</td>
<td>1000</td>
</tr>
<tr>
<td>Figure</td>
<td>1 x 250</td>
<td>250</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>7072</td>
</tr>
</tbody>
</table>

August 29, 2014
ABSTRACT

Carsharing is an innovative transportation mobility solution which offers the benefits of a personal vehicle without the burden of ownership. Free-floating carsharing service a relatively new concept and is gaining popularity because it offers additional flexibility allowing one-way auto rental and charging users usage by minute. Traditionally carsharing services require returning the rented vehicle to the same location where rented with a minimum rental duration. Since free-floating service is a very new addition in the overall transportation system, the empirical research is still very limited. This study focuses on identifying the impact of land-use variables on free-floating carsharing vehicle rental choice and parking duration of Car2Go services in Austin, Texas on a typical weekday between 9:00 AM to 12:00 PM. Two different methodological approaches, a namely logistic regression model approach and a duration model technique are used for this purpose. The results of this study indicate that land-use level demographic variables, the carsharing parking policy, and numbers of transit stops effect the usage of free-floating carsharing vehicles.

Key Words: Free-floating carsharing, Land-use, Carsharing Rental Choice, Duration Model
1. INTRODUCTION
The urban development in the United States is highly dependent upon the automobile based mobility system. This automobile dependency has led to many transportation and environmental problems including traffic congestion, greenhouse gas emissions, air and noise pollution, and foreign oil dependency (Kortum and Machemehl, 2012). In addition, personal vehicle ownership costs its owners more than $9,000 per year (AAA, 2013). Carsharing programs might be viewed as a novel substitute to personal vehicle ownership in urban areas. Carsharing is an innovative short-term car rental option that provides the mobility of using a private car without the burden of car ownership costs and responsibilities (Shaheen et al., 1998). Such programs facilitate reductions in household vehicle ownership by motivating road users’ behavior towards personal transportation decisions as mobility on demand as opposed to an owned asset. The carsharing concept is recognized as a sustainable mobility solution because recent research studies have shown its positive environmental impact at least in three-ways (Firnkorn and Müller, 2011). First, a reduction of total carbon dioxide emissions (Firnkorn and Müller, 2011; Haefeli et al., 2006); second, reduction in household vehicle holdings (Shaheen and Cohen, 2012); third, reduction of vehicle miles traveled (Martin et al., 2010).

Carsharing service providers allow members on-demand vehicle rental where the basic concept is “pay-as-you-drive”. Members of the carsharing program have access to a fleet of vehicles in the service network and pay per use. Currently two different forms of carsharing programs are in-practice: 1) traditional/station-based carsharing systems; and 2) free-floating/one-way carsharing systems. Traditional carsharing programs (such as ZipCar) offer short-term rental generally with an hourly pricing option and require users to return vehicles to the original location of renting. On the other hand, free-floating carsharing programs (such as Car2Go and DriveNow) allow users one-way car rental where cars can be rented and dropped-off at any location within a specified service area. The main advantage of free-floating carsharing programs over traditional station-based carsharing programs is its flexibility because it overcomes the limitation of traditional carsharing system where the rented vehicle requires dropping off to the same station where it was rented.

German carmaker Daimler’s Car2Go program was the first major initiative allowing users one-way rental within a city’s operating area. In 2008 Car2Go was first launched in Ulm, Germany and later expanded its service in 28 cities in 8 different countries across Europe and North America. Recently German carmaker company BMW is also offering one-way free-floating carsharing program DriveNow in five cities in Germany and one city in the U.S. In the U.S. Car2Go was first launched in Austin, Texas in November, 2009 as a pilot carsharing project for city of Austin employees and later the service was opened to the general public in May, 2010 (Kortum and Machemehl, 2012). Presently, over 16,000 Car2Go members use 300 identical Car2Go vehicles in Austin and all vehicles can be parked for free at any City of Austin or State of Texas controlled meter, parking space or designated parking space for Car2Go vehicles within its operating area (Car2Go, 2014; Greater Austin 2013 Economic Development Guide, 2013).

Research studies on carsharing vehicles have routinely claimed that there is an association between land-use development patterns and the use of carsharing vehicles. However, there is no rigorous attempt to explain the causal mechanism that generates the association between land-use and demand for carsharing vehicles. For example, Car2Go vehicles in different cities exercise free on-street parking at any meter in their operating area. Therefore it is worthwhile to investigate if the parking cost has any effect on carsharing vehicle use. Again, easy access to alternative transportation modes has been emphasized in the
carsharing literature is likely to affect the use of carsharing vehicles. However, no studies exist in the literature that investigated such relationships for free-floating carsharing programs empirically. Most of the earlier studies investigated station-based car-sharing programs, however, research on free-floating sharing programs is very limited. One of the factors that limit such investigation is the availability of data about the use of free-floating carsharing vehicles.

This study focuses on identifying impacts of land-use variables on free-floating carsharing usage, namely Car2Go vehicle rentals and parking duration (duration unused). Carsharing vehicle rental and unused durations are likely to depend on various land-use characteristics such as parking cost, auto-ownership, dominant age of the population in the location, employment density, transit facilities, and land-use mix diversity level. We use a logistic regression model approach to identify factors that affect Car2Go vehicle rentals and a hazard-based duration model approach to identify factors that affect their unused duration in Austin, Texas on a typical weekday between 9:00 AM to 12:00 PM. Survey data obtained from actual carsharing members’ can be helpful to understand those factors; however, the costs associated with such surveys are non-trivial. In this study real-time free-floating Car2Go vehicle location and condition data over time (in five minute intervals) are obtained from a Car2Go application programming interface (Car2Go API) at no cost. The data provides robust information on vehicle rentals, movement, and availability across the Austin, TX area. Since Car2Go uses remote control technologies to activate and monitor vehicle locations, the data collection infrastructure is already integrated within the system and can be accessed by the public in real time. This is the primary set of data that enables one to identify the usage of free-floating carsharing vehicles. The vehicle location dataset is supplemented by aggregate level land-use and transit-stop data obtained from the Capital Area Metropolitan Planning Organization (CAMPO) and Capital Metro. A parking survey conducted by CAMPO for the Austin area is also used as a data source for the study.

The remainder of this paper is organized as follows. The next section presents a brief review of the literature on the topic of this paper. The third section presents the modeling methodology, while the fourth section presents a description of the data set used. The fifth section presents model estimation results and the sixth section offers concluding thoughts.

2. BACKGROUND AND LITERATURE REVIEW

2.1 History of Carsharing

Originating in Switzerland during the mid nineteenth century, the carsharing program has been popular in Europe for decades. During the early 1980s, this concept was first implemented in the US as pilot research projects to evaluate its feasibility (see Fricker and Cochran, 1982; Crain & Associates, 1984). Commercial operation of station-based carsharing was first established in 1998 in Portland, Oregon (Katzev, 2003). Carsharing service has evolved from initial neighborhood residential services where shared-use vehicles parked in designated areas within a neighborhood to robust more-targeted focus customers in spatially denser areas having easier access to alternative transportation modes (Shaheen et al., 2009). This sector has been emerging and growing over time, and the number of U.S. users has grown from 12,000 in 2002 to over 890,000 in January of 2013 (Shaheen and Cohen, 2013).

A relatively new approach of carsharing services is the free-floating operation which is unique because its flexibility over traditional services. Traditional services offer a short-term round-trip rental
option that require the user to return vehicles to the original location of renting, whereas free-floating carsharing services allow users one-way rental within a specified service boundary that facilitates discretionary activities. Free-floating carsharing service may be more appealing to consumers because it is free from fixed costs such as minimum rental duration, a booking fee, and a minimum monthly usage associated with traditional carsharing services as it only charges users usage by minutes. Free-floating carsharing vehicles are described as being close to private automobiles in terms of convenience because they permit one-way rentals. The service convenience is actually inferior to the personal automobile because there is out of vehicle travel time associated with accessing the mode. The convenience of such services may increase as autonomous free-floating carsharing vehicles (or autonomous Taxis) may be available to consumers.

The free-floating carsharing service Car2Go was introduced first in Ulm, Germany by Daimler Auto Group. Commercial operation of Car2Go service in the US was first established by the same company in 2010 in Austin, Texas. The fleet size and the operating area of the service were expanded over time. As of April, 2014 a total of 300 identical Smart for Two vehicles provide access to over 16,000 Car2Go members in the Austin area. Real-time information is provided to the consumers about available vehicle location, internal and external condition, and fuel level or electric charge level from internet, hotline or a Car2Go Smartphone application.

2.2 Carsharing Vehicle Usages

In consumer research carsharing is viewed as an access-based consumer product. Historically access-based consumption was considered as an inferior consumption alternative over ownership because of the attachment associated with owned belongings (Bradhi and Eckhardt, 2012). However, the confluence of shifting demographics and changing dynamics in the economy facilitated carsharing to gain popularity due to perceived environmental benefits and cost savings (Bradhi and Eckhardt, 2012; Shaheen et al., 2010). In fact, ownership is often perceived as burdensome while the short-term access mode allows flexibility and economic savings (Bradhi et al., 2012). Technological advancement, increased use of Smart phones and availability of real-time route information also help road users to reconsider their travel mode choice alternatives including carsharing.

From the transportation and land-use perspective, carsharing is considered a service providing improved mobility for travelers without a vehicle of their own, who use transit, share rides, or move on bicycle or by foot, but still require access to a personal vehicle for a trip segment. It is an access-based consumer service characterized as “pay-as-you-drive” and often an economical alternative for individuals who may not require daily auto access and drive less than 10,000 miles per year (Shaheen et al., 2010). It is also believed that college and university students will benefit from the service because they can gain access to a fleet of vehicles without bearing the burden of ownership (Shaheen et al., 2004). Therefore it is observed that carsharing programs in the US, both traditional and free-floating, have developed close relationships with universities around the country (Kortum and Machemehl, 2011). Recent research on the demographics of carsharing members characterized them as mostly young professionals, single and age between 21 to 38 years (Bradhi and Eckhardt, 2012).

A number of US cities supporting smart growth recognize the environmental and social benefit of carsharing and have developed parking policies allocating on-street parking spots for carsharing vehicles (please see Shaheen et al., 2010 for a good review about the carsharing parking policy in US cities). For example, in Austin, Car2Go vehicles can be parked on-street at any City of Austin or State of Texas
controlled meter or pay station parking space free of charge (car2go Austin Parking FAQs, 2014). Such a carsharing friendly parking policy is likely to have positive effects on the usage of carsharing vehicles especially in university towns having parking challenges.

Overall, it can be seen that there are a host of land-use level socio-economic and demographic factors, as well as parking policies, and attitudinal variables that can affect the usage of carsharing vehicles. The number of studies investigating the effect of land-use variables on carsharing vehicle usage is very limited because most of the studies are based on station-based carsharing services. Moreover, despite the recognition that parking cost and transit service can affect the usage of carsharing services, most of the earlier studies on free-floating carsharing did not consider their effects on carsharing vehicle usage. In this study, we contribute to the literature on carsharing vehicles by identifying the impact of land-use variables on free-floating carsharing vehicle rental choice and parking duration (vehicle unused duration). Two different methodological approaches are used to investigate free-floating carsharing vehicle usage. The first approach is a logistic regression model approach where the dependent variable is a binary outcome that indicates whether or not renting of an available carsharing vehicle occurred; and the second is a duration model approach where the dependent variable is a continuous variable representing the unused duration of carsharing vehicles fleet time.

3. METHODOLOGY

Two different methodologies are adopted to identify land-use factors affecting the usage of Car2Go carsharing vehicles. The first approach is a logistic regression model approach where the dependent variable is a binary outcome representing the choice of rental of an available free-floating vehicle within a given time period. This approach will be helpful to identify land-use factors that affect the usage of available carsharing vehicles in a three hour time window. The second approach is a duration model where each available vehicle observed at a specific time period is observed for three hours or until it is rented again. The methodologies are presented in the following subsections.

3.1 Logistic Regression Model

In the logistic regression model, the response variable \((y_j)\) is binary or dichotomous in nature where \(y_j\) takes the value of 1 for a given Car2Go available vehicle \(j\) if the vehicle becomes unavailable (because of renting) during the time of observation and takes the value of 0 if the vehicle remains available to rent at the end of observation period. The equation for the standard logistic regression model is:

\[
y_j' = x_j \beta + \xi_j, \quad y_j = 1 \text{ if } y_j' > 0
\]

Where, \(y_j'\) corresponds to the latent propensity for vehicle \(j\), \(x_j\) is a vector of explanatory variables, \(\beta\) is a corresponding vector of coefficients that will be estimated and \(\xi_j\) is an idiosyncratic random error term assumed to be identically and independently standard logistic distributed. In the latent variable model, we do not observe the latent propensity \(y_j'\), however, we observe in the sample data whether an available Car2Go vehicle is still available or not at the end of an observation period. The probability that Car2Go vehicle \(j\) would become unavailable (rented) is:
\[ \text{Prob}(y_j = 1 | \mathbf{x}_j) = \text{Prob}(y_j^* > 0 | \mathbf{x}_j) = G(\mathbf{x}_j \beta) \]

Where, \( G(.) \) is the cumulative density function for the error term which is assumed to be logistically distributed.

\[ G(\mathbf{x}_j \beta) = \frac{e^{\mathbf{x}_j \beta}}{1 + e^{\mathbf{x}_j \beta}} \]

Similarly, the probability of remaining available (not rented) is:

\[ \text{Prob}(y_j = 0 | \mathbf{x}_j) = \text{Prob}(y_j^* < 0 | \mathbf{x}_j) = G(-\mathbf{x}_j \beta) = 1 - G(\mathbf{x}_j \beta) \]

To estimate the model by maximum likelihood, we use the likelihood function for each \( j \). The parameters to be estimated in the logistic regression model are the \( \beta \) parameters.

### 3.2 Duration Model

We let \( T \geq 0 \) be a continuous random variable that denotes the time measured in minutes until an available Car2Go vehicle becomes unavailable (rented) again. Then \( t \) denotes a particular value of \( T \) and the cumulative distribution function (cdf) of \( T \) is defined as

\[ F(t) = P(T \leq t), t \geq 0 \]

The instantaneous renting rate per unit of time is \( \lambda(t) \), commonly referred to as the instantaneous hazard rate in duration model literature. The mathematical definition for the hazard in terms of probabilities is

\[ \lambda(t) = \lim_{h \to 0^+} \frac{P(t \leq T, t + h | T \geq t)}{h} \]

The hazard function can be easily expressed in terms of the density and cdf very simply.

\[ \lambda(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)} = \frac{dF/dt}{S(t)} = \frac{-dS/dt}{S(t)} = -\frac{d \ln S(t)}{dt} \]

where \( S(t) \) is referred as a “survivor function” in duration literature and in reliability literature it is referred to as a “reliability function”. In this study the authors prefer the term “availability function”.

The shape of the hazard function (instantaneous renting rate) has important implications for duration analysis. Two parametric shapes for instantaneous renting rate \( \lambda(t) \) are considered here. In the first case, the instantaneous renting rate is assumed to be constant implying that there is no duration dependence or duration dynamics, mathematically, \( \lambda(t) = \lambda, \forall t \geq 0 \). The conditional probability of being rented does not depend on the elapsed time since it has become available to rent. The constant-hazard assumption corresponds to an exponential distribution for the duration distribution. The instantaneous renting rate function with covariates is: \( \lambda(t; \mathbf{x}) = \exp(\mathbf{x} \beta) \), where \( \mathbf{x} \) is a vector of explanatory variables and \( \beta \) is a corresponding vector of coefficients to be estimated.

In the second case, the instantaneous renting rate function is generalized to accommodate duration dependency resulting in a Weibull distribution for the duration data. The hazard rate in this case allows...
for monotonically increasing or decreasing duration dependence and is given by \( \lambda(t) = \gamma \alpha t^{\alpha-1}, \gamma > 0, \alpha > 0 \). The instantaneous renting rate function with covariates is: 
\[
\lambda(t; x) = \exp(x^\beta) \alpha t^{\alpha-1},
\]
where \( x \) is a vector of explanatory variables and \( \beta \) is a corresponding vector of coefficients to be estimated. When \( \alpha = 1 \), the Weibull distribution reduces to the exponential with \( \lambda = \gamma \), implying no duration dependence. If \( \alpha > 1 \), there is a positive duration dependence and the instantaneous renting rate is monotonically increasing. If \( \alpha < 1 \), there is a negative duration dependence and the instantaneous renting rate is monotonically decreasing.

4. DATA DESCRIPTION

4.1 Data sources

Five different data sources are used in this study. The first dataset, as discussed earlier is the real-time 24 hours Car2Go vehicle location and condition data in five minute intervals obtained from Car2Go. The data provides robust information on vehicle rentals, movement, and availability across the Austin, TX area. This primary data enables one to identify the usage of free-floating carsharing vehicles. The second dataset is the transportation analysis zone (TAZ) level land-use data provided by CAMPO for year 2005. The CAMPO data provided a host of land-use level socio-demographic information including population, number of households, household size, 2005 median household income, autos owned, total employment etc. The third dataset is land-use level demographic data based on 2010 census obtained from the Capital Area Council of Governments (CAPCOG). Total population, race/ethnicity, as well as other items, at block level are obtained from the CAPCOG data source. The fourth dataset is a parking survey conducted in year 2010 provided by CAMPO that encompasses the study area. The fifth dataset describes transit stops of the study area as an open data source created by Capital Metro, the public transit service provider.

4.2 Formulation of Dependent Variables/ Car2Go Vehicle Usage Data

Car2Go vehicle location data was collected for a typical weekday (Tuesday) on February 25th, 2014 from 12:00 AM to 12:00 PM in five minute intervals. The data collection effort was automated by setting up a Windows Task Scheduler operation that collected and stored data thereby reducing possibilities of human error. Car2Go API data provides information about available vehicles only. Therefore, if a vehicle is observed at time \( t \) and after some time \( h \) the vehicle is no longer found in the dataset, then it implies that the vehicle was rented within the \( t \) and \( t+h \) time intervals. In this study, each of Car2Go vehicles observed available at 9:00 AM is tracked every five minutes until 12:00 PM to see if they were rented. For instance if one Car2Go vehicle (say, vehicle 1) is available at 9:00 AM and 9:05 AM but it does not appear in the dataset at 9:10 AM, then it is assumed that the vehicle became unavailable at 9:10 AM because it was rented sometime between 9:05 AM and 9:10 AM. The availability status (rented vs. not rented) of each observed vehicle is used as the depended variable of the logistic regression model.
The dependent variable for the duration model analysis is continuous. If a vehicle was found to become available at 7:00 AM and then was rented again at 9:10 AM, then the total unused duration is 130 minutes\(^1\). The availability status and the duration unused (available duration prior to next rental) of each vehicle was recorded for further analysis.

The characterization of Car2Go vehicles unused duration can be represented by Figure 1. Vehicle 1 corresponds to the sample observation described earlier. The dataset also contains observations represented by vehicle 2, 3, and 4, for which the beginning or ending times are synonymous with the sampling beginning or ending times. Such observations are referred to as right or left censored observations. The duration models used in this study are capable of accounting for the right censored observations. Vehicle type 3 illustrates the case for which the beginning time when it became available is unknown and is referred to as a left censored observation. Vehicle type 4 represents those cases having both left and right censoring. By following the history of the vehicles represented by types 3 and 4 one

\(^1\) Actually the unused duration for the vehicle is somewhat different because the vehicle observed available at 7:00 AM can become available anytime between 6:55 AM to 7:00 AM. Similarly, vehicle could be rented anytime between 9:05 AM to 9:10 AM. Therefore the actual unused duration is a range from 125 minute to 135 minute. Kortum and Machemehl (2011) found the median rental duration varies between 10 to 12 minutes and therefore observation in five minutes interval was assumed reasonable for the analysis that also simplifies data handling.
can trace the start time of the unused duration. However, this paper focuses only on the unused duration in a given weekday and therefore limits the analysis to midnight to noon on a typical weekday.

4.3 Data Assembly

The final sample was assembled in a number of steps. First, a total of 240 available vehicles observed at 9:00 AM are mapped to one of the 172 TAZs within the Car2Go service boundary using geographic information system (GIS). This operation appended vehicle availability status and unused duration information with the TAZ level land use data for each vehicle. Second, land-use level demographic data obtained from CAPCOG are aggregated into the TAZ level. Third, the number of transit stops located in each TAZ is calculated using the GIS platform. Fourth, parking survey data is also aggregated into TAZ level in order to identify TAZs that impose parking charges. Finally, TAZ level aggregated data on demographics, number of transit stops, and parking cost are appended with the Car2Go vehicle location data.

4.4 Sample Characteristics

A total of 240 Car2Go vehicles were observed in the 172 TAZs encompassing the Car2Go operating area in Austin, TX at 9:00 AM on February 25th, 2014. At the time of observation the 240 available vehicles were observed in 100 TAZs. The CAMPO data divides all TAZs into five area types, namely, CBD, Urban Intense, Urban, Suburban and Rural. The Car2Go operating area comprises the first four area types. Total number of zones in each area type, distribution of the vehicle locations in the four area types, average population and employment in the corresponding area types are presented in Table 1. Although higher numbers of available vehicles are observed in the area types with higher population (i.e. urban intense and urban area types), the average number of available vehicles per zone is highest in the CBD area.

<table>
<thead>
<tr>
<th>Area Type</th>
<th>Number of Zones</th>
<th>Number (%) of Available Vehicles</th>
<th>Avg. No of Vehicles Per TAZ</th>
<th>2005 Population</th>
<th>2005 Employment</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBD</td>
<td>11</td>
<td>37 (15.4%)</td>
<td>3.36</td>
<td>117</td>
<td>3826</td>
</tr>
<tr>
<td>Urban Intense</td>
<td>37</td>
<td>102 (42.5%)</td>
<td>2.76</td>
<td>2059</td>
<td>2267</td>
</tr>
<tr>
<td>Urban</td>
<td>46</td>
<td>87 (36.3%)</td>
<td>1.89</td>
<td>1660</td>
<td>720</td>
</tr>
<tr>
<td>Suburban</td>
<td>6</td>
<td>14 (5.8%)</td>
<td>2.33</td>
<td>1193</td>
<td>360</td>
</tr>
</tbody>
</table>

Summary statistics of land-use variables for the Car2Go service area weighted by the number of available cars are presented in Table 2. As one can see from Table 2, the locations of Car2Go vehicles at 9:00 AM within the service area shows on average more than 11% of the households have no car. Again, locations of available vehicles are mostly in neighborhoods where the mean percentage of adult population is about 87%. This is not surprising because the central part of the city including the downtown area and the neighborhoods (CBD and Urban Intense area Types) around the University of Texas at Austin have higher concentrations of Car2Go members (Kortum and Machemehl, 2011). The average household size is 2.11 which is slightly lower than the 2005 average of 2.40 in Austin (see City of Austin, 2009).
Table 2 Summary Statistics of the Land Use Variables Used in the Models

<table>
<thead>
<tr>
<th>Land Use Variables</th>
<th>Mean</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent of Household having no car (Year 2010)</td>
<td>11.01</td>
<td>0.87</td>
<td>46.75</td>
</tr>
<tr>
<td>Percent of Over 18 Population (Year 2010)</td>
<td>86.91</td>
<td>65.65</td>
<td>100.00</td>
</tr>
<tr>
<td>Average Household Size (Year 2005)</td>
<td>2.11</td>
<td>0.00</td>
<td>3.66</td>
</tr>
<tr>
<td>Number of Transit Stops (Year 2010)</td>
<td>11.02</td>
<td>1.00</td>
<td>31.00</td>
</tr>
<tr>
<td>Employment Density (Year 2005:# of total employment/acre)</td>
<td>24.35</td>
<td>0.07</td>
<td>199.47</td>
</tr>
</tbody>
</table>

All 100 TAZs are divided into two categories based on the median income of the TAZs. A total of 33 (14%) vehicles are located in high income TAZs (Income>$60,000). Based on parking policy, all 100 TAZs are divided into two groups and only seven are found to charge parking fees. A total of 29 (12%) of available vehicles were observed in TAZs where parking is not free.

A total of 110 (46%) of the observed vehicles were rented between 9:00AM-12:00 PM. The total unused duration of the rented vehicles ranges from 5 minutes to 705 minutes with an average unused duration of 331 minutes. The unused duration of the 130 unrented vehicles during the period of observation (12:00 AM to 12:00 PM) ranges from 180 minutes to 720 minutes with an average of 480 minutes.

5. MODEL ESTIMATION RESULTS

This section presents a discussion of the logistic regression model estimation results for carsharing vehicle rental choice and the duration model estimation results for carsharing vehicle unused duration (Table 3 and Table 4 respectively).

5.1 Logistic Regression Model

A number of land-use level socio-demographic attributes affect the choice of whether an available vehicle will be rented or not. As expected, the propensity of an available Car2Go vehicle to be rented increases as the percentage of households having no car increases. The availability of free-floating carsharing vehicles increases the mobility options for households having no car and therefore increasing the availability of vehicles in those areas increases the likelihood of those vehicles being rented. Households having their own car may perceive the walk-time to access a free-floating vehicle more burdensome compared to those having no personal automobile choice. As the percentage of over 18 population increases in a TAZ, the propensity to rent an available Car2Go vehicle from that TAZ also increases. Car2Go vehicles can accommodate only two adults in the vehicle and therefore may be more attractive where concentrations of adult individuals is relatively higher (as opposed to suburban areas where single family homes with young children are predominant). Again, as the number of transit stops increases in a TAZ, the propensity to rent an available Car2Go vehicle from that TAZ increases. Vehicles located in high median income TAZs (Income>$60,000) are more likely to be rented compared to vehicles located in low-income TAZs. Higher income may be working as a proxy variable representing a higher level of education and environmental consciousness. Individuals with such behavioral attitudes are likely to consciously make an effort to have a smaller carbon footprint on the environment and chose an environmental friendly free-floating vehicle as a modal option. Increasing household size in a TAZ increases the propensity of the
vehicles located in those TAZs to be rented. On the other hand, as the employment density increases in a TAZ, the propensity to rent an available Car2Go vehicle from that TAZ decreases. This result is not surprising because free-floating carsharing vehicles are less likely to be used on a daily basis to access employment opportunities; rather they are likely to be used for discretionary activities justifying the sign of the employment density variable.

Table 3: Logistic Regression Model Result of Rental Choice

<table>
<thead>
<tr>
<th>Explanatory Variable</th>
<th>Parameter Estimate</th>
<th>z-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-12.07</td>
<td>3.32</td>
</tr>
<tr>
<td>TAZ Land-Use Measures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socio-Demographics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent of Household having no car (x0.1)</td>
<td>0.71</td>
<td>3.01</td>
</tr>
<tr>
<td>Percent of Over 18 Population (x0.1)</td>
<td>1.02</td>
<td>2.96</td>
</tr>
<tr>
<td>High-Income TAZ (vs. Low-Income TAZ)</td>
<td>1.48</td>
<td>2.46</td>
</tr>
<tr>
<td>Average Household Size</td>
<td>0.84</td>
<td>2.10</td>
</tr>
<tr>
<td>Employment Density</td>
<td>-0.03</td>
<td>3.31</td>
</tr>
<tr>
<td>Parking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parking Cost (vs. doesn't have Parking Cost)</td>
<td>2.39</td>
<td>2.50</td>
</tr>
<tr>
<td>Transit Availability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Transit Stops</td>
<td>0.05</td>
<td>2.39</td>
</tr>
<tr>
<td>log-likelihood Value</td>
<td>-141.15</td>
<td></td>
</tr>
</tbody>
</table>

As expected, there is a higher renting propensity of a Car2Go vehicle in a TAZ with parking charges compared to a TAZ with free parking. This finding has important policy implications in that parking policy directly affects the usage of such services. Again, as the number of transit stops increases, the propensity to rent an available Car2Go vehicle from that TAZ increases. This validates the assumption of carsharing services in areas with good transit services making intermodal trips easier.

The CAMPO area type definition is also used as an explanatory variable in the carsharing vehicle rental choice model, however, the effect was found to be non-significant. Perhaps the disaggregated level land-use variables included in the model explain the data variability more precisely than the aggregated land-use classification.

5.2 Duration Model Results

Table 4 shows the estimated covariate effects for the two parametric duration model specifications. The estimated value of the alpha parameter is $1.33>1.0$ for the Weibull parametric distribution of hazard function (instantaneous renting rate) and is statistically significant at the 0.05 level. This alpha parameter $>1$ indicates that the hazard (the probability to be rented) is increasing monotonically over time and therefore the results of Weibull parametric distribution of hazard function is appropriate for the data.

The effect of land-use level socio-demographic characteristics indicates that locations with higher percentages of households having no car have a higher hazard (i.e., a smaller unused duration) than locations with lower percentages of households having no car, probably because of increased mobility options availability to these locations. A 1% increase in 0-car households in a TAZ increases the
probability to be rented of the vehicle(s) located in that TAZ by 5.2%. Vehicles located in areas with higher percentages of adult population (age>18 years) have a higher hazard (i.e., a smaller unused duration). A 1% increase in the percentage of adult population increases the renting probability of a vehicle located in that location by 2.9%.

Table 4 Free-floating Carsharing Vehicle Unused Duration Model Results

<table>
<thead>
<tr>
<th>Explanatory Variable</th>
<th>Weibull Distribution</th>
<th>Exponential Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Parameter Estimate</td>
<td>z-stat</td>
</tr>
<tr>
<td>Constant</td>
<td>-12.53</td>
<td>-8.61</td>
</tr>
<tr>
<td>TAZ Land-Use Measures</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Socio-Demographics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent of Household having no car (x0.1)</td>
<td>0.52</td>
<td>4.41</td>
</tr>
<tr>
<td>Percent of Over 18 Population (x0.1)</td>
<td>0.29</td>
<td>2.03</td>
</tr>
<tr>
<td>Parking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parking Cost (vs. doesn't have Parking Cost)</td>
<td>1.58</td>
<td>4.05</td>
</tr>
<tr>
<td>Transit Availability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Transit Stops</td>
<td>0.03</td>
<td>2.22</td>
</tr>
<tr>
<td>Time Dependency Parameter</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alpha</td>
<td>1.33</td>
<td>3.02</td>
</tr>
<tr>
<td>log-likelihood Value</td>
<td>-238.71</td>
<td></td>
</tr>
</tbody>
</table>

As expected, parking cost has a very significant impact on the usage of carsharing vehicles (Millard-Ball et al., 2005; Kortum and Machemehl, 2011)). At any point in time, locations that charge parking fees have about 100*[exp (1.58)-1] =385% greater hazard (probability of an available vehicle being rented) than the locations where parking is free-of charge. Since Car2Go vehicles exercises free on-street parking, those vehicles are likely to be chosen for making trips in those locations to gain the benefit of free-parking. Again, vehicles available in those areas where parking is challenging are likely to be rented because of higher trip production and attraction rates in those areas.

The effect of transit indicates that locations with greater numbers of transit stops have a higher hazard (i.e., a smaller unused duration) than locations with smaller numbers of transit stops, probably because of increased intermodal connectivity allowed by better transit availability. An additional transit stop in a TAZ increases the hazard (probability to be rented) of the vehicle(s) located in that TAZ by 3%.

6. CONCLUSION

Free-floating carsharing service is gaining popularity because it permits users one-way auto rental and charges users by the minute. Such additional flexibility over traditional station based carsharing services allows users to select this mode to perform discretionary activities. This study focuses on identifying the impact of land-use variables on free-floating carsharing vehicle rental choice and parking duration of Car2Go services in Austin, Texas on a typical weekday between 9:00 AM to 12:00 PM. Two different methodological approaches, namely a logistic regression model approach and a duration model are used for this purpose.
The results of this paper identify the importance of land-use level socio-demographic attributes, support from local governments by facilitating carsharing parking policy and the availability of transit facilitating intermodal transportation. It appears that the usage of free-floating carsharing vehicles is higher in neighborhoods having lower auto-ownership, greater numbers of adult population (age >18 years), larger household size, and higher household income levels. The study results confirm the impact of these factors with observed real-time data.

Support from local government plays a vital role that promotes carsharing. The result of this study indicate that parking policy in favor of carsharing services plays a very important role in carsharing vehicle usage. Since carsharing vehicle usage has significant environmental benefit, supporting these services with additional designated parking spots may help increase usage, and thereby reduce travel by personal automobile. Better transit service availability also positively affects the usage of carsharing vehicles.

The vehicle usage data employed in the study was collected without cost from the Car2Go API. The land-use data are typical of information maintained by metropolitan planning organizations for travel demand models and can also be accessible in most cases. Often the expense in data collection imposes difficulty in research. The innovative dataset used in this research can shed light on alternative available data sources in transportation planning research.

However, the data used in this study does not allow investigation of trip purposes of the rented vehicles. Disaggregate level usage data from a survey could help characterize trip purposes.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Dr. Daniel Yang of the Capital Area Metropolitan Planning Organization (CAMPO) in Austin, Texas for his help with the CAMPO land use and parking data used in the analysis. Brice Nichols of Puget Sound Regional Council provided the python code to download the Car2Go API Data. The authors are grateful to Lisa Smith for her help in formatting this document.
REFERENCES

American Automobile Association (AAA). 2013. Your Driving Costs 2012. Available at, 


\cmtagis1\gis\Download_Folder\cmta_stops.shp

Car2Go Austin Parking FAQs, 2014. Available at, 
https://www.car2go.com/common/data/locations/usa/austin/Austin_Parking_FAQ.pdf

City of Austin, 2009. Community Inventory Demographics: Demographic & Household Trends. 
Retrieved from, ftp://ftp.ci.austin.tx.us/GIS-

Urban Mass Transportation Administration, Washington, D.C.

Firnkorn, J, Müller, M., 2011. What will be the environmental effects of new free-floating carsharing 

Fricker, J.D, Cochran, J.K., 1982. A queuing demand model to optimize economic policy-making for a 
shared fleet enterprise. Presented at the October ORSA/TIMS Conference in San Diego, CA.

Greater Austin 2013 Economic Development Guide, 2013. Car2Go, Available at, 
http://www.businessintexas.org/austin-economic-infrastructure/car2go/

the Environment, Transport, Energy and Communications, Bern.


Prediction, Mode Share, and Vehicle Allocation Optimization Methodologies. Report 
SWUTC/12/476660-00079-1. Project 476660-00079.

Where and How It Succeeds. Transportation Research Board of the National Academies, 
Washington, D.C.


Big urban probe data for the provision of advanced traveler information services and traffic management schemes

Dr. Josep-Maria Salanova Grau
Research Associate
Centre for Research and Technology Hellas – Hellenic Institute of Transport, Thessaloniki, Greece
jose@certh.gr

Dr. Evangelos Mitsakis
Associate Researcher
Centre for Research and Technology Hellas – Hellenic Institute of Transport, Thessaloniki, Greece
emit@certh.gr

Iraklis Stamos
Research Associate
Centre for Research and Technology Hellas – Hellenic Institute of Transport, Thessaloniki, Greece
stamos@certh.gr

Dr. Georgia Aifandopoulou
Research Director
Centre for Research and Technology Hellas – Hellenic Institute of Transport, Thessaloniki, Greece
gea@certh.gr
Abstract
This paper presents a framework for data collection, filtering and fusion, together with a set of operational tools in an effort to validate, analyze, utilize and highlight the added-value of probe data. An application of the proposed framework is presented for the city of Thessaloniki, Greece, where two types of datasets are considered: probe data and traffic data collected through conventional methods. The probe dataset is comprised of individual objects’ pulses (smart devices, navigators, etc.) tracked throughout the network at constant and pre-defined locations (“stationary” probe data collection) or during the whole trip of an “object” that continuously generates pulses (“dynamic” probe data collection). The conventionally collected traffic datasets originate from inductive loops, cameras and radars. The main difficulty when processing probe data derived from predefined locations is to identify the correct set of detections for estimating path travel times, especially in dense urban networks. In such cases, experience and literature suggest that three main error sources exist: a) co-existent modes of transport (private vehicles, buses, pedestrians, bicycles etc.), b) existence of more than one possible paths between two detectors and c) existence of stops or trips ending between two detectors. In the same direction, a difficulty for processing the data of the dynamic network is that of map-matching, which is the problem of identifying the path of a vehicle through a link-based representation of the road network, based on a sequence of time-stamped oriented coordinates. The paper concludes with operational applications of the datasets under real world conditions in the city of Thessaloniki, highlighting key parameters, advantages and limitations of the use of probe data in urban areas.

Keywords: mobility data, floating car data, probe data, big data

Introduction
Technological advances have been lately attributed with an increased quality and quantity of mobility-related data. People, being nowadays in a constant state of information sharing, have transformed into active players of the data collection process. Luring them with the provision of real-time services in a variety of fields, ranging from routing suggestions to traffic conditions updates, technological advances have managed to engage people in a constant exchange of information. Smart phones and other portable devices that offer “internet on the go” have succeeded in overcoming the naturally-set temporal and geographical limitations, making it possible for users to be connected at any place and time, thus rendering them to online information transmitters by sending and receiving valuable information to/from the
content and services providers. Yet, the challenge of producing the best possible end-products out of these big datasets is twofold; on the one hand there is a need for developing algorithms able to fuse, filter, validate and process big amounts of data (almost) at real-time, while on the other hand, there is a constant need for developing new applications and services for providing innovative and advanced traveler information services and traffic management schemes based on these data and processing capabilities.

The structure of this paper is as follows. The first chapter discusses big data sources for the transportation and mobility sector, while the second chapter presents an applied case for the city of Thessaloniki. Chapter three is dedicated to the different uses of big data and presents a series of applications to real networks of Thessaloniki, Greece and Barcelona, Spain. The paper concludes in chapter four with recommendations and suggestions for future research efforts and directions.

**Mobility data sources**

**Conventional traffic data sources**

Conventional data sources mostly aim at collecting aggregated data through sensors installed at fixed locations of the network. The measurements concern vehicles’ traffic flow, speed and traffic lanes’ occupancy. Various sensors and technologies have been developed for counting and classifying vehicles and for measuring speed and occupancy. Sherry L. S. (2001) lists the limitations of the most representative traditional traffic measurements sensors and technologies, concluding in that the common limitations are related to installation and calibration costs, low coverage and low performance under adverse weather conditions. Regarding the same sensors and technologies, Ledug G. (2008) presents the lifetime and costs (purchase, installation, operation and management), concluding that intrusive sensors have a lifetime of 5 years and high maintenance costs, while non-intrusive sensors have a lifetime of 10 years and high purchase and installation costs.

**Probe data**

Probe data is a type of crowd-sourced data collected from individuals, including vehicles, passengers, travelers or pedestrians. In this case, data is aggregated after the collection phase, which significantly increases the quality of the collected data and multiplies the capabilities for processing this data and having a better representation of the mobility patterns in a city. The sensors for probe data are mostly owned by individuals themselves. Probe data can be classified into stationary and floating. The stationary probe data is collected at fixed locations,
while floating probe data is collected throughout the whole network. Stationary probe data is collected at various points located along the network by detecting communication protocol identities, such as Bluetooth, or through automatic plate number recognition (APNR) systems. The significant drawbacks of APNR systems are the high costs, calibration needs and the low performance under high traffic flow conditions or adverse weather conditions. Respectively, the major drawback of using Bluetooth detectors is the need for data processing and the penetration rate of the technology, which has however significantly increased during the last years (70% of all new vehicles will have Bluetooth connectivity in 2016 (Strategy Analytics, 2010), while the number of mobile phones and portable devices with Bluetooth will exceed 600 million in 2015 (IMS Research, 2012)).

Floating car data and floating passenger data is continuously generated by moving objects (cars or users) equipped with a smart device able to calculate its own location through GPS or A-GPS. It was initially used (and still is) by fleet operators aiming at monitoring the location of the vehicles and tracking routes, events and incidents along the vehicles’ routes. The major drawback of FCD is related to the accuracy of the measurements, which is actually solved by map-matching the position of the vehicle to a static map. Map-matching is used for identifying the path of a vehicle through a link-based representation of the road network. The input consists of a sequence of time-stamped oriented coordinates, which generate the trajectory data and a link-based representation of the road network, henceforth map. The output of the map-matching can be the collection of instantaneous position and speed of a vehicle or a sequence of positions. There are two significant issues that perplex the map-matching process: the GPS accompanied measurement error and the frequency of the tracking itself. The GPS accuracy ranges from 3 to several decades of meters depending on the quality of the GPS receiver and the presence of obstacles. The tracking frequency is a fundamental issue and should be carefully examined. High frequencies imply larger quantities of data, which are associated with higher computational efforts and monetary costs but better trajectory tracking. Low frequencies imply worst trajectory tracking but also a reduction in the processing requirements and the operational costs. Figure 1 depicts examples of FCD and the map-matching procedure.
As stated by Chawathe (2007), the magnitude of the errors is often larger than the distance between two link features. The author states that the 90% confidence region surrounding a track point may encompass several link features, which means that the use of two consecutive points for estimating travelled distance or speed may have a significant error, especially in dense urban areas. In order to solve this, various map matching methodologies have been developed that can be classified in those based on geometry, which do not take into account the topological constraints of the road network and in those based on topology, which take them into account. Topology methods are based on network consistency, where the road network has topological constraints that may be respected (e.g. the route cannot shift from a highway to a local road if there is no physical exit ramp). Both methods are very resource-intensive, especially when backtracking techniques are applied for re-drawing the vehicle’s path backwards, which means that a set of GPS points (a segment) is matched at each instant instead of a unique GPS point. Most map-matching techniques are based on a combination of the two approaches for jointly matching sets of tracked points.

### Data from social media

Social media is the most recent mobility-related data source and the one with the highest potential (up to 2000000 twits per hour (Leetaru et al. 2013)), but at the same time the most difficult dataset to extract and analyze information-rich content, due to its format (free text) and its lack of geo-reference (only 38% of the twits are geo-referenced (Leetaru et al. 2013)). Semantic methodologies for mining and fusion are being developed for analyzing social media content, using pre-determined keywords.
**Mobility data in Thessaloniki**

Thessaloniki is the second largest city in Greece, with a total of more than 1 million citizens in its greater area, covering a total of 1500 km² with an average density of 665 inhabitants per km². The total number of vehicles in the city exceeds 777,544, including private cars, heavy vehicles and motorcycles. A complete description of the mobility indicators of the city of Thessaloniki can be found in Mitsakis et al. (2013a). Three mobility data sources are used in Thessaloniki for collecting mobility-related data: conventional sensors (loops, radars and cameras), probe data (stationary and floating) and social content.

**Conventional data sources in Thessaloniki**

There are three sets of conventional mobility data sensors in Thessaloniki:

- The surveillance system of the Peripheral Ring Road, monitoring a total of more than 100,000 vehicles per day in both directions (green circles in Figure 2).
- Thessaloniki’s Urban Mobility Management System installed in the city center, monitoring more than 50,000 vehicles per day (red diamond in Figure 2).
- The traffic lights management system of the wider metropolitan area of the city (blue triangles in Figure 2).

![Figure 2 – Mobility sensors in the city of Thessaloniki](image)
Stationary probe data

The Bluetooth detectors network of the city of Thessaloniki is comprised of 43 roadside devices, installed at selected intersections throughout the road network of the city, as shown in Figure 2.

![Figure 3 – Bluetooth detectors network in the city of Thessaloniki](image)

More than 100000 Bluetooth-equipped devices are detected every day, generating a total of 300000 detections at the 43 locations, which can be classified into 1750 different groups of MAC identities from 400 different trademarks. Figure 4 shows the distribution of detections between the different groups of MAC identities, where it can be observed that 90% of the detections are related to less than 20% of the groups of MAC identities, and less than 10% of the trademarks.
Floating car data

The network of moving sensors (Floating Car Data) is comprised of more than 1200 taxi vehicles, circulating in average between 16 and 24 hours per day, which periodically (each 6 seconds) send pulses containing their location and speed. The total amount of data collected and processed reaches 2500 pulses per minute, with daily totals at approximately 1.5 million. Figure 5 shows the concentration of pulses per minute during 14 days.

The quality of FCD will be significantly enriched with the provision of cooperative mobility services (connected vehicles and infrastructures) to more than 250 taxis. These vehicles will be able to provide data related to their position and speed per second as well as to detect congestion.

Social Media

The data obtained from social media content is related to individual Facebook check-ins in various locations of the city, which are obtained at real-time. A total of 1500 locations in the
city center account for more than 35000 daily check-ins of Facebook users. Figure 6 below shows the accumulated check-ins and the check-ins during the last hour in the historical part of the city center.

![Figure 6 – Concentration of check-in in the city center](image)

**Mobility Data processes**

The data collected by sensors and systems described above are used for estimating mobility-related indicators of the city, such as travel time estimation along the main routes of the city, traffic flow estimation, traffic congestion detection and other mobility related characteristics. These new datasets can be used for providing both higher quality mobility services and for applying new and more effective mobility management schemes.

New methodologies are being developed for the estimation of mobility patterns and Origin-Destination matrices, road hazard detection and calibration of route choice models as well as macroscopic and microscopic multimodal traffic simulation models.

**Traffic flow measurements from stationary probe data in Thessaloniki**

Traffic flow can be estimated from detections if the penetration level of Bluetooth-equipped devices is representative of the traffic flow. In order to investigate this, traffic flow has been measured in one of the intersections equipped with a Bluetooth detector device, in order to evaluate the relation between the detections and the actual traffic flow, as shown in Figure 7.
Figure 7 – Relation between detected Bluetooth devices and measured traffic flow without filtering the detections

Figure 8 shows that there is a strong relation, which can account for 37% of the traffic flow being detected by Bluetooth detectors.

After processing and filtering the Bluetooth detections and deleting double detections during an interval of time, which are related to not moving objects, the corrected relation is approximately in the order of 20%, as shown in Figure 9.
Figure 7 – Relation between detected Bluetooth devices and measured traffic flow after filtering the detections at one intersection

Table 1 shows the relation between detections and traffic flow for each of the applied filters. The best fit is obtained with the 15 minutes filter, with a relation of 20%.

<table>
<thead>
<tr>
<th></th>
<th>Without filtering</th>
<th>5' filter</th>
<th>15' filter</th>
<th>60' filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relation</td>
<td>0,3412</td>
<td>0,2179</td>
<td>0,1972</td>
<td>0,0442</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0,9166</td>
<td>0,9193</td>
<td>0,9337</td>
<td>0,8594</td>
</tr>
<tr>
<td></td>
<td>-26% / 75%</td>
<td>-23% / 61%</td>
<td>-22% / 57%</td>
<td>-35% / 79%</td>
</tr>
</tbody>
</table>

Travel time estimation from stationary probe data in Thessaloniki

Individual travel times can be estimated by detecting the same MAC identity at various locations of the network, thus average travel time can be obtained by aggregating individuals travel times. The Bluetooth detected devices are used for estimating travel time by matching them to the Bluetooth detectors network locations. Individual travel times are filtered, in order to eliminate outliers and noise, such as en-route stops, trip ends between detectors or detours deviating from the pre-defined paths or records from other transportation modes, such as pedestrians, bicycles, users of the public transport system as well as atypical vehicles such
as couriers or delivery vehicles. Details on this methodology can be found in Mitsakis et al. (2013b).

More than 600000 individual trips are matched per week by the network of 43 Bluetooth detectors of Thessaloniki. Figure 10 shows two estimators for the travel time for one route during the whole day, as well as the individual travel time matched by the system.

![Figure 8 – Travel time estimators from individual travel times](image)

**Real time traffic conditions estimation from probe data**

By merging all the data presented above, traffic congestion in the city of Thessaloniki is estimated at real-time and forecasted in a short-term basis. Both travel time measured at selected routes by the network of Bluetooth detectors and instantaneous speeds measured from the FCD are converted into traffic flow by using volume-delay functions. All traffic flows are merged into a non-linear mathematical program by means of a modification of the Data Expansion algorithm presented by Lederman and Winter (2009). Finally, all measurements are forecasted by an Auto Regressive model for 15, 30, 45 and 60 minutes and the process is repeated, in order to obtain traffic flows for the forecasted scenarios. Figure 11 shows the entire computational process.
Conclusions and further work
The amount of mobility-related data is huge. Technology allows for shifting from aggregated to disaggregated data collection at user level. This results to a need for aggregating methodologies and more post-processing needs, but it has enormously increased the quality of the collected data and therefore the quality of the provided mobility services, especially in urban areas.

Data curators have become an important agent in the mobility theater, passing from an era with an important lack of data to the era of data, where data privacy and quality are receiving most of the attention.

In order to foster this data ecosystem, open data schemes should be promoted by public authorities and institutions, which should make big mobility datasets available to both the research community and service providers.
References


Identifying Driving Risk Factors to Support Usage-Based Insurance using Smartphone Personalized Driving Data

Yi-Chang Chiu¹, Yu-Luen Ma², Xianbiao Hu³*

¹ Associate Professor, Department of Civil Engineering and Engineering Mechanics, University of Arizona, Tucson, AZ, USA, chiu@email.arizona.edu
² Professor, Katie School of Insurance and Financial Services, Illinois State University, Normal, IL, USA, yma@ilstu.edu
³* Director of R&D, Metropia, Inc., Tucson, AZ, USA. xb.hu@metropia.com

Words: 5908
Figures and Tables: 16
ABSTRACT

Historically, automobile insurance premiums in the U.S. were largely based on various static factors, such as socio-demographic and geographical information, without taking individualized risk factors into consideration. Advances in the development of modern GPS devices have led some insurers to start looking into variations of Usage Based Insurance (UBI) or Pay-As-You-Drive-And-You-Save (PAYDAYS) research. Progressive® Insurance Company, for example, monitors the driving behavior of their policyholders by installing a GPS device in their cars for a certain period of time, offering adjusted premiums based on the GPS reading, although shortcomings with this approach have been observed.

A comprehensive research framework and approach to analyzing vehicle use and/or driver behavior data to advance knowledge about driving exposure factors that are closely linked to crash risk is presented in this paper. The detection criteria for crash/near crash event will first be identified and carefully calibrated. Such criteria will be applied to the Metropia Mobile user database to automatically detect all crash/near crash events associated with Metropia Mobile users. The driving behavior evaluation model will then be built to quantitatively evaluate each user’s driving behavior based on the available data including GPS trajectory, traffic speed, roadway geometry and crash event information. In addition, through exploring partnership with insurance companies, loss severity information for the detected crash event will be retrieved.

The results of our study will further existing knowledge about driving exposure factors that are closely linked to crash risk. The manner in which these factors affect loss frequency and claim severity will also help insurers re-structure their existing pricing models to allow for variation in premiums based on these characteristics, and provide the actuarial foundation for advanced forms of PAYDAYS insurance pricing.

Keywords: Usage Based Insurance (UBI), Pay-As-You-Drive-And-You-Save (PAYDAYS), Driving Risk Factors, Personalized Driving Data, Smartphone data collection, insurance pricing
1. INTRODUCTION

Historically, automobile insurance premiums in the U.S. were largely based on various static factors, such as socio-demographic and geographical information, without taking individualized risk factors into consideration. Today, most people’s driving patterns are not incorporated into actuarial pricing by insurers simply because such data has not been available, even though driving habits, traffic congestion and vehicle dynamics can all be potential hazards that lead to deadly accidents.

The emergence and subsequent rapid advances with new information and communication technologies (ICT) such as GPS devices, cellphone, Bluetooth, etc., now offer the capability of collecting high-fidelity and high-resolution travel data in a cost-effective manner. These technologies, especially smartphones and mobile application platforms, also permit continuous data collection related to driver behavior, so long as the smartphone device remains in operation. Given the ever growing cellular phone market,\(^1\) it has now become much easier to track and understand traveler activity, and travel patterns.

Researches of applying the latest ICT to collect data and evaluate driving behavior started to show up in the last decade. A recent effort in the individual driving behavior is the 100-car naturalistic driving study [1]. Drivers are monitored and recognized as unsafe, moderate safe and safe according to frequencies of crashes/near-crashes. The results indicate that hard braking, inattention, and tailgating are the top three at-risk behaviors among drivers. Unsafe drivers are more likely to engage in the at-risk behaviors and decelerate/swerve greater than the safe drivers. The results also imply that improper braking and inappropriate speeds are positively related to crash/near-crashes. Different traffic and weather conditions are also studied separately for the driving behavior and crash risk. The unsafe drivers drive more aggressively regardless of traffic conditions.

Another example of using new information and communication technologies in the driver behavior evaluation can be found in [2], who proposed a framework for profiling drivers by at-risk behavior using driving pattern, spatial and temporal characteristics and driver characteristics. Second-by-second GPS data observations are collected from 106 drivers in Sydney over several weeks. Behavioral measures are summarized as maximum, average, minimum and standard

\(^1\) A whopping 90% of adults in the United States own cellular phones, of which 65% own a smartphone.
deviation of speed, acceleration and deceleration, distance at 75% of speed limit or over speed, number of sharp celeration ($\geq 4 \text{ m/s}^2$), etc. Also, in the last decade the Data Acquisition System (DAS) or In-Vehicle Data Recorders (IVDR) have been introduced to collect detailed driving behavior data, but the usage was limited due to the high hardware cost, and the research sample size are usually insufficient [3, 4]. More examples of applying ICT to evaluate driving behavior can be found in [5-8]. Those researches, however, are mostly limited in the model validations, and the amount of users in the experiments are usually low.

Advances in the development of modern GPS devices have led some insurers to start looking into variations of Usage Based Insurance (UBI) or Pay-As-You-Drive-And-You-Save (PAYDAYS) research. Progressive® Insurance Company, for example, monitors the driving behavior of their policyholders by installing a GPS device in their cars for certain period of time, offering adjusted premiums based on the GPS reading. There are two shortcomings with this approach. First, due to the high hardware cost, this short-term tracking does not permit the insurer to continuously monitor and track policyholders’ driver behavior, which provides little incentive for drivers to maintain safe driving habits once the trial period ends. Additionally, the data collected by Progressive’s and other insurers’ systems depict only the speed and acceleration/deceleration of the vehicle. Other important risk factors such as location information (e.g. an intersection of higher risk than a mid-link) and traffic flow are not traceable by the GPS devices.

In this 24 month research project titled “Pay-As-You-Drive-And-You-Save (PAYDAYS) Insurance Actuarial Study”, we propose to answer the following major research question: How can we collect and use various types of data, including vehicle trajectories, associated link traffic volume and speed data, time of day information, and desired claim data, to more accurately identify driving exposure factors that are closely linked to crash risk, and quantify the relationship between crash, hazard, driving behavior and congestion levels, as well as accident claims? In order to answer this question, a statistical analysis model pertaining to the technical narratives on our collected data as well as research methodology plan will be developed. The primary emphases include actuarial analysis based on vehicle trajectories, dynamic traffic data

---

2 This research is supported by Federal Highway Administration Broad Agency Announcement “Pay-As-You-Drive-And-You-Save (PAYDAYS) Insurance Actuarial Study” project. Contract award number DTFH61-13-C-00033.
and hazard proxy variables with the goal of establishing the linkage between driving risks and resulting claims.

In this paper, the overall research framework to be used in this research project will be presented, which mainly includes data collection, general research work flow and actuarial analysis research approach. Current progress -- including crash detection threshold calibration and preliminary trajectory analysis -- will also be presented.

The rest of this paper will be organized as follows: Section 2 presents the methodology being developed for this research project, including the overall research work flow, data collection which covers the list of data being collected in this research, and data collection mechanism, and the actuarial analysis research approach to identify the driving risk factors. Section 3 reviews the research progress to date, including the hard brake experiment performed earlier this year to better calibrate the crash detection threshold, and the preliminary analysis on the GPS trajectory data. Section 4 concludes this research and discusses future research directions.

2. RESEARCH METHODOLOGY

2.1. OVERALL RESEARCH WORK FLOW

In this section, a comprehensive research framework and approach to analyzing vehicle use and/or driver behavior data to advance knowledge about driving exposure factors that are closely linked to crash risk is presented. We envision the research to span over a 24-month period, with the main modules in this research to include the following:

- Data collection: specifying what kind of raw data we are collecting, and how we collect these raw data
- Data processing: describing how different kinds of raw data will be processed in this research to extract useful information for further actuarial analysis purposes
- Actuarial analysis: explaining how the data will be used in this research for actuarial analysis purposes, what model will be built to analyze the data to identify the driving risk factors, etc.

After the GPS trajectory is sent back to the cloud server, those data will be stored and processed for PAYDAYS research purposes. The detection criteria for crash/near crash event will first be identified and carefully calibrated. Such criteria will be applied to the Metropia Mobile user database to automatically detect all crash/near crash events associated with users. Such crash
detection module developed will be used to detect the crash event in a real time fashion, which means if abnormal situations are observed from the GPS trajectory sent back from user’s smartphone and certain criteria is met, the system will consider the chance of this user running into crash to be very high. Certain subsequent modules will be triggered to verify the crash event, query the person injury and property loss information from users, and retrieve the claim cost from partner insurance company, if available.

The driving behavior evaluation model will then be built to quantitatively evaluate each user’s driving behavior based on the available data, including GPS trajectory, traffic condition, roadway geometry and crash event information. In addition, through exploring partnership with insurance companies, loss severity information for the detected crash event will be retrieved.

The results of our study will further existing knowledge about driving exposure factors that are closely linked to crash risk. The manner in which these factors affect loss frequency and claim severity will also help insurers restructure their existing pricing models to allow for variation in premiums based on these characteristics, and provide the actuarial foundation for advanced forms of PAYDAYS insurance pricing.

All of those data collected, together with the insurance claim data, will be used for further statistical and actuarial analysis to identify the risk factors associated with the driving behavior, and quantify their relationship with the auto accidents. The overall research work flow for this project can be found in Figure 1 below.
The main procedure is as follows:

Step 0: Determine the criteria for auto crash detection. This is an important module since accurately defining the logic and calibrating the threshold is critical to efficiently identify the potential crash event. A near crash detection experiment was performed earlier this year to achieve this goal.

Step 1: Collect GPS trajectory from user’s smartphone, and such data will be sent back to the cloud server in a real-time fashion.
Step 2: Through the real time trajectory processing engine, the system will identify the potential crash event based on the comparison between user information and the pre-defined crash detection rules. Data cleaning will happen in this step prior to the crash event detection, a systematic method will be used to identify and remove invalid data points.

Step 3: If potential crash event has been identified, an email with pre-defined questionnaire will be automatically sent to the user, the purpose of this email is to confirm the accident event.

Step 4: After receiving the confirmation of crash event, if the user is also a customer of our partner insurance company, the claim cost will be retrieved when it becomes available for further research analysis purpose. Otherwise, a second email with questionnaire will be sent to the user querying the estimated claim cost based on the person injury and property loss information.

Step 5: Actuarial analysis with statistics modeling to reveal the driving risk factors and establish the linkage between driving risks and resulting claims.

The system will continue steps 1 through 4 until it collects sufficient data samples for statistics and actuarial analysis purpose. Section 2.4 below will further discuss the actuarial analysis research approach (Step 5 above) for this PAYDAYS project.

2.2. DATA COLLECTION

The dynamic new smartphone-based data collection mechanism opens up the possibilities of allowing long-term continual data collection that identifies true high risk behaviors. As one of the critical steps of this project, data collection provides the foundations for the different modules in the system research framework, including potential crash event identification, driving risk factors identification, actuarial analysis to quantify the relationships hazards and accident claims and so on.

In this research, in order to answer the question of what are the factors that are attributable to auto accidents, and how can we use various types of data to more accurately quantify the relationship between driving, traffic and auto accidents, we propose to collect various risk factor data, as well as indicating whether an accident resulted from a particular journey. The primary data are categorized into:

1) Vehicle trajectory data
2) Roadway geometry
3) Time-varying traffic dynamic
4) Claim data

In this study we use Metropia Mobile, a recently developed smartphone-based app designed to improve mobility management, to keep track of people’s driving behavior, surrounding traffic, and other trajectory data to develop models that investigate the relationships between various driving hazards and insurance claims. Metropia Mobile is a recently available mobile traffic app that uses prediction and coordinating technology combined with user rewards to incentivize drivers to cooperate, balance traffic load on the network, and reduce traffic congestion.

When a user starts a trip, the internal GPS module is activated and starts to record the second-by-second latitude/longitude data location and instantaneous moving speed. These data allow detailed position, velocity, acceleration and deceleration details to be stored and analyzed. One unique feature of the Metropia data is that its backend server also estimates traffic speed and volume for each link that the vehicle traverses. Such data -- combining both user trajectory and link speed/volume information -- has rarely been seen in prior research, permitting a unique opportunity to link critical traffic congestion factors that lead to certain driving behaviors and crash potential. For example, if a driver exhibits stop-and-go or abrupt accelerate/decelerate behavior, it is usually difficult to tell if this is simply due to the driver’s behavior, or because of heavy traffic conditions. When both driving and traffic data are linked together, one can discern hazards caused by driving behavior from those caused by congestion levels.

2.2.1. VEHICLE TRAJECTORY DATA

Detailed trajectory data and driving behavior data for each trip are collected during the trip validation process. When a user starts a trip with the app, the internal GPS module built in the smartphone is activated and starts to record the second-by-second latitude/longitude data location and instantaneous moving speed. These data, including detailed position such as latitude, longitude and altitude, heading, velocity, acceleration and deceleration will be collected at fine time interval and sent back to the cloud server, where they will be stored and used for further analysis both in the online or offline fashion.

The main attributes of the trajectory data can be found below:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UserID</td>
<td>The unique ID of the user</td>
</tr>
<tr>
<td>Attribute</td>
<td>Description</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>TripID</td>
<td>The unique ID of the particular trip</td>
</tr>
<tr>
<td>Latitude</td>
<td>The latitude of this particular GPS point</td>
</tr>
<tr>
<td>Longitude</td>
<td>The longitude of this particular GPS point</td>
</tr>
<tr>
<td>Altitude</td>
<td>The altitude of this particular GPS point</td>
</tr>
<tr>
<td>Heading</td>
<td>The heading of vehicle at this moment</td>
</tr>
<tr>
<td>Speed</td>
<td>The instantaneous travel speed of the vehicle</td>
</tr>
<tr>
<td>Acceleration</td>
<td>The instantaneous acceleration value of the vehicle</td>
</tr>
<tr>
<td>Timestamp</td>
<td>The timestamp of this particular GPS point</td>
</tr>
<tr>
<td>LinkID</td>
<td>The ID of the road segment that user is currently driving on</td>
</tr>
</tbody>
</table>

2.2.2. **ROADWAY GEOMETRY**

The geographic network stored on cloud server with the main purpose of providing routing and navigation functionalities to the users. A typical traffic network usually includes a set of links and nodes, i.e. $G = (N, A)$ where $N$ is the set of nodes $\{1, 2, \ldots, n\}$ and $A$ the set of link. The key attributes relevant with this research is the latitude and longitude of the nodes, the link types, i.e. whether a road segment is located on freeway, arterial, local streets, the speed limit of the road segment, the number of lanes of the road segment and so on.

Table 2 describes the list of data attributes for the geographic network dataset.

**Table 2. Roadway geometry data**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinkID</td>
<td>The ID of the road segment that user is currently driving on</td>
</tr>
<tr>
<td>Speed Limit</td>
<td>The speed limit of the current road segment</td>
</tr>
<tr>
<td>Facility Type</td>
<td>The type of links, such as freeway, arterial, local street, ramp</td>
</tr>
<tr>
<td>Number of lanes</td>
<td>Number of lanes of the road segment that user is currently driving on</td>
</tr>
<tr>
<td>Link length</td>
<td>The length of the road segment</td>
</tr>
<tr>
<td>Start nodeID</td>
<td>The ID of the start node for this particular road segment</td>
</tr>
<tr>
<td>Start latitude</td>
<td>The latitude of the start node</td>
</tr>
</tbody>
</table>
2.2.3. **TIME-VARYING TRAFFIC DYNAMIC**

On the cloud server, the time-varying traffic dynamic, i.e. the traffic dataset including time-dependent traffic speed, traffic flow, etc., with the time interval of five minute at each traversed link is hosted and keeps updating every 5 minutes. The main purpose of these data is to feed into the routing engine and guide users to the desired destination while avoiding the traffic congestion. The traffic dataset of interest mainly includes time-dependent link speed and volume information. List of attributes of the traffic data can be found in Table 3 below.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinkID</td>
<td>The unique ID of a road segment</td>
</tr>
<tr>
<td>Date</td>
<td>The date of the traffic data</td>
</tr>
<tr>
<td>Timestamp</td>
<td>The timestamp of this particular traffic data</td>
</tr>
<tr>
<td>Link average speed</td>
<td>The instantaneous average speed of this road segment</td>
</tr>
<tr>
<td>Link volume</td>
<td>The instantaneous traffic volume of the current road segment</td>
</tr>
</tbody>
</table>

2.2.4. **CLAIM DATA**

In this research, we consider two possible scenarios – with or without insurance company’s claim data. In the scenario in which at least one insurance company is willing to partner with the research team in this project and is willing to share claim data as they come in, when an incident occurs and a claim is filed, if the policyholder is a Metropia traffic app user, this claim will be cross-referenced with the corresponding reservation and trajectory record. Claim data for the participant who is not insured with a partnering insurance company and is confirmed to have been involved in an accident will be collected through questionnaire, which is automatically sent out to the driver sometime after the accident happens.
2.3. **DATA PROCESSING**

2.3.1. **DATA CLEANING**

Data cleaning process will be applied prior to the data being used in the research, which is very important to ensure the quality of system input and accurate analysis result. A systematic method has been developed to identify and remove invalid data points. Some typical invalid data scenarios include but are not limited to:

- GPS being inaccurate under certain scenarios, such as when it’s around tall buildings or indoors. In a dense city packed with skyscrapers like Los Angeles and New York, finding an unobstructed patch of sky to divine your location can be a serious challenge.
- GPS being inaccurate due to network issues. There’s a component of assisted GPS to the system, which requires a cell signal and numerous nearby towers with which the phone can triangulate its location. If the cellphone network connection is weak, for example under 3G or 2G network, the location result may not be as accurate as with strong signal.
- GPS point “jumps” when the speed of vehicle moving is low. It is sometimes observed that when the vehicle driving speed is low, there’s a chance that the heading, location of the GPS will not be very accurate.
- Issues with the data transmission. During the data transmission from cell phone to the cloud server, due to some known or unknown issue, the data may be lost or sent over multiple times.
- Network data quality issue, such as missing speed limit for the road segment.

2.3.2. **DATA JOINS AND SAMPLE DATA**

For the actuarial analysis purpose, if we perform data joins between the multi-source data mentioned in Section 2.2, in the end for actuarial analysis purpose, for each GPS point, we will have the following attributes:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>UserID</td>
<td>The unique ID of the user</td>
</tr>
<tr>
<td>TripID</td>
<td>The unique ID of the particular trip</td>
</tr>
<tr>
<td>Latitude</td>
<td>The latitude of this particular GPS point</td>
</tr>
<tr>
<td>Longitude</td>
<td>The longitude of this particular GPS point</td>
</tr>
<tr>
<td>Altitude</td>
<td>The altitude of this particular GPS point</td>
</tr>
<tr>
<td>Heading</td>
<td>The heading of vehicle at this moment</td>
</tr>
<tr>
<td>Speed</td>
<td>The instantaneous travel speed of the vehicle</td>
</tr>
<tr>
<td>Acceleration</td>
<td>The instantaneous acceleration value of the vehicle</td>
</tr>
<tr>
<td>Timestamp</td>
<td>The timestamp of this particular GPS point</td>
</tr>
<tr>
<td>LinkID</td>
<td>The ID of the road segment that user is currently driving on</td>
</tr>
<tr>
<td>Distance to Downstream Intersection</td>
<td>Distance to the next intersection that drivers need to turn</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>--------------------------------------------------------</td>
</tr>
<tr>
<td>Next Movement</td>
<td>The next turning movement at the intersection, such as turn right, turn left, U turn</td>
</tr>
<tr>
<td>Facility Type</td>
<td>The type of links, such as freeway, arterial, local street, ramp</td>
</tr>
<tr>
<td>Number of lanes</td>
<td>Number of lanes of the road segment that user is currently driving on</td>
</tr>
<tr>
<td>Link average speed</td>
<td>The instantaneous average speed of the current road segment</td>
</tr>
<tr>
<td>Link volume</td>
<td>The instantaneous traffic volume of the current road segment</td>
</tr>
<tr>
<td>Speed limit</td>
<td>The speed limit of the current road segment</td>
</tr>
</tbody>
</table>

Most attributes in the above table are pretty straightforward and could be obtained by simple data joins. One thing to mention is some special data processing is needed for the “Next Movement” and “Distance to downstream intersection”, which requires the algorithm to scan through the user’s whole trip trajectory, extract the decision points where drivers need to make the turns, and compute these two values for each GPS point.

2.4. ACTUARIAL ANALYSIS RESEARCH APPROACH

The general research work flow in Figure 1 described the systematic framework of how the GPS trajectory data will be collected and applied to identify crash event, how the crash event data and claim cost data will both be obtained and correlated to facilitate the actuarial analysis. In this section, the overview of PAYDAYS auto insurance actuarial analysis research approach will be presented. The emphasis of this actuarial analysis will be placed on several aspects including risk measurement design, statistical modeling and data analysis.

2.4.1. ACTUARIAL ANALYSIS RESEARCH APPROACH FOCUSES

To achieve the objectives of this insurance actuarial study, the actuarial analysis research approach proposed in this project will involve three main steps.

We will first define and explain how we identify the driving risk factors and develop a baseline analysis that describes the characteristics of our sample. The risk factors that we believe will affect driving behavior and may potentially lead to accidents will firstly be defined. The characteristics of those risk factors, such as distribution of driver accelerations, decelerations will then be analyzed with the sample data from the database.

Secondly, a correlation analysis of risk measurements that are likely to be linked with accidents will be performed, the purpose of this analysis is to reveal the relationship between different risk
factors, validate various assumptions on the risk measurements that may or may not be contributing to the accidents, and provide theoretical foundations to the full statistical analysis later.

For the full statistical analysis which is the third step, generalized linear models (GLMs) will be utilized for predicting both loss frequencies and loss severities and provide implications for insurance rate making. GLM is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value. By removing the restrictive assumptions usually seen in the linear models, GLM eliminates the chances of the models being built violate real world practice, and it’s extensively used in the statistics.

2.4.2. **KEY VARIABLES DESCRIPTION**

We envision the key variables of measuring driving hazards to include the variables that can be observed from the user GPS trajectory or can be derived from joining trajectory data with the geometric network as well as dynamic traffic data. For example, since driving speed and acceleration/deceleration can be obtained from the second-by-second detailed GPS trajectory, by associating the GPS point data with the network road segment speed limit information, it can be determined whether the driver drove at a speed higher than the speed limit. In additional, it can also be inferred how many times the driver has made harsh brake during the trip by looking at the deceleration values.

For each trip, the following key variables can be defined and calculated:

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speeding</td>
<td>if driver is driving at a speed higher than speed limit of the road segment by a certain threshold</td>
</tr>
<tr>
<td>Relative Speed</td>
<td>if the driver’s driving speed deviates from other drivers on the same road at the same time for more than a certain threshold</td>
</tr>
<tr>
<td>Braking</td>
<td>if deceleration is lower than given threshold</td>
</tr>
<tr>
<td><strong>Time in traffic</strong></td>
<td>the time when traveling in traffic and driving slowly</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Acceleration</strong></td>
<td>if acceleration is higher than given threshold</td>
</tr>
<tr>
<td><strong>Peak Time</strong></td>
<td>the time when driving during peak hours</td>
</tr>
<tr>
<td><strong>Late time</strong></td>
<td>time when driving at midnight or early morning</td>
</tr>
<tr>
<td><strong>Left Turn</strong></td>
<td>the number of times driver makes a left turn during the trip</td>
</tr>
<tr>
<td><strong>Right Turn</strong></td>
<td>the number of times driver makes a right turn during the trip</td>
</tr>
<tr>
<td><strong>Mileage</strong></td>
<td>the total distance of the trip</td>
</tr>
<tr>
<td><strong>Travel time</strong></td>
<td>the total length of time traveled during this trip</td>
</tr>
<tr>
<td><strong>Accident</strong></td>
<td>a Boolean variable, the value is true if this trip is associated with an accident, otherwise false.</td>
</tr>
<tr>
<td><strong>Crash Loss</strong></td>
<td>the claim amount, if this trip is associated with a crash event</td>
</tr>
</tbody>
</table>

3. **CURRENT PROGRESS**

In this section, current progress of this PAYDAYS research project will be presented, which mainly includes the hard brake experiment to calibrate the crash detection module, and a preliminary analysis on the GPS trajectory data.

3.1. **HARD BRAKE EXPERIMENT**

The capability to automatically detect an accident is one of the key requirements to this PAYDAYS research. As shown in the overall research work flow in Figure 1, the activities to verify accidents either via questionnaires or through an insurer partner will be triggered only after a crash event has been detected.

The GPS modules in the smartphone is constantly updated in fine time intervals and send the data to the trajectory database on the cloud server, and these acceleration/deceleration values can be computed in a real-time fashion. Combined with other supplemental information, such as speed afterwards and location changes, a harsh deceleration can be detected and flagged as a potential crash/near crash event if it exceeds a certain threshold. In this design, the detection is primarily based on scanning the speed-related data in the recorded vehicle trajectories and identifying the hard breaking/large deceleration occurring in most of the crash/near crash...
In the past, we’ve relied on literature to define this type of threshold value; however, such a value may not be applicable for our purpose because of the differing measurement instruments used in literature.

The experiment conducted earlier this year is specifically for the PAYDAYS research study. We collected a set of field data from a controlled experiment that involved actual hard breaking behavior by different drivers using various types of vehicles. These data will help determine appropriate deceleration thresholds for the detection of potential crash/near crash events by the system.

3.1.1. **EXPERIMENT SETUP**

Time: Weekend (Saturday) morning

Location: Shopping mall parking lot (El Con Mall, Tucson, Arizona)

Conditions: Clear, dry and warm

Total number of vehicles: 5 (4 sedans + 1 SUV)

Total number of phones: 7 (4 android phones + 3 iPhones)

Equipment:

1. A light-weight baseball training net
2. Paint bucket
3. Tripod and camera/camcorder

Procedure:

1. Place the baseball net at a fixed location on the test track.
2. Measure 50-100 feet upstream and place the bucket on the right side of the track.
3. Ask the driver to reach a speed higher than 30 mph.
4. The driver cannot break until the vehicle passes the bucket.
5. The driver needs to break hard enough to avoid hitting the net.

A video showing an example of the procedure for this test can be viewed at http://youtu.be/uwr8YaV0kXc

---

3 It is possible for a crash to occur without a hard break if the driver is not cognitively capable of processing dangerous driving conditions or perform a timely reaction (e.g. driving under influence), or if the weather condition is icy and a crash occurs as a result of vehicle skidding. These types of cases won’t be detected by the proposed approach. Such events, however, account for a relatively small percentage of crash incidents when compared to those with normal driving reaction.
Figure 2. Experiment location overview
Figure 3. Hard Break Field Testing photos from a test conducted in Tucson, Ariz.

3.1.2. **COLLECTED DATA**

A total of 78 samples were collected in the experiment. Below is a brief visualization of the testing procedure and speed fluctuations. The maps of Figure 4 and Figure 5 show vehicle trajectory as collected by phone, where it can be seen that the vehicle drove back and forth during the experiment period. The graphs of the two figures show the vehicle was first speeding up, and then the speed quickly goes back to 0 as the testers brake very hard (X axis – time, Y axis - speed)

Figure 4. Data collected by Phone 1 (upper - GPS trajectory, lower – speed fluctuations)
3.1.3. PRELIMINARY RESULTS

The deceleration values of the hard brakes during the experiment were plotted in the historical diagram below. As can be seen, the deceleration data exhibits a bell shaped distribution with just a few observations of larger deceleration values.

![Deceleration distribution](image)

Figure 6. Hard Brake Deceleration Histogram

A more detailed statistical results analysis can be found in Table 6 below. The mean value of deceleration/ft. per square second is about -21.1, the median is -19.7, and the standard deviation is 7.36. At confidence level of 95%, Z value is 1.681, and confidence interval is [-22.5, -19.7].

<table>
<thead>
<tr>
<th>Deceleration/ft. per square sec.</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>-21.09672219</td>
</tr>
</tbody>
</table>
3.2. Preliminary trajectory analysis

In this section, sample user trajectory data is extracted from the user database and examined to better understand user driving behavior, including user’s instantaneous speed fluctuation, the frequency of hard brake and/or jackrabbit starts, the acceleration and deceleration values, etc. We present some brief examples below.

Figure 7 through Figure 9 visualize the GPS trajectory of three users, with each including the time-dependent speed fluctuation and celeration (acceleration is the number is positive, and deceleration if negative) changes. A preliminary analysis reveals that User 3 braked not only the most frequently, but also the hardest, so it’s reasonable to assume he/she was driving more aggressively than the other two users.

41 Some of the deceleration values were observed to be as high as -40 ft/s², which according to our preliminary analysis was due to the Anti-lock Braking System (ABS).
Figure 7. User 1 morning trip analysis (speed fluctuation and celeration changes)

Figure 8. User 2 late night trip analysis (speed fluctuation and celeration changes)
The GPS trajectory is also correlated with the road network geometry information to better understand users’ driving behavior and how deceleration behavior/action relates to roadway information. In Figure 10, the red signifies the driver stopping or driving at relatively low speeds, and green indicates the user driving at higher speeds. The graphic allows us to easily determine that the traffic was light and the red dots indicate slow movement mostly at intersections.
Figure 10. Driving speed fluctuations along the travel trajectory

The case studies in this section demonstrated the feasibility of applying the data collected by smartphone to quantitatively evaluate the driving behavior. All the trajectory data collected will be further analyzed in this research by correlating with other data source including roadway geometry, time-varying traffic dynamics and claim data. Statistical models will be built to performed to reveal the relationship between different risk factors, validate various assumptions on the risk measurements, and in the end, to predict both loss frequencies and loss severities and provide implications for insurance rate making.

4. CONCLUSIONS AND FUTURE RESEARCH

Advances in the development of modern GPS devices and smartphone technologies have made the application of Usage Based Insurance (UBI) or Pay-As-You-Drive-And-You-Save (PAYDAYS) possible, which has been demonstrating significant advantages over the traditional insurance pricing model. A comprehensive research framework and actuarial analysis research
approach to analyzing vehicle use and/or driver behavior data to advance knowledge about driving exposure factors that are closely linked to crash risk is presented in this paper. The next steps in this 24 month research project include defining and computing risk measurements, performing the correlation analysis between different risk factors, and building actuarial analysis model to quantify the relationships hazards and accident claims. We believe that as time goes, more personalized driving data will become available and used in the statistical model for analysis, the research team can further move this research forward, so that in the end, the results of our study will further existing knowledge about driving exposure factors that are closely linked to crash risk. The manner in which these factors affect loss frequency and claim severity will also help insurers re-structure their existing pricing models to allow for variation in premiums based on these characteristics, and provide the actuarial foundation for advanced forms of PAYDAYS insurance pricing.

5. REFERENCES


7 Ericsson, E. 2000. Driving pattern in urban areas-descriptive analysis and initial prediction model. Lund Institute of Technology, Department of Technology and Society, Traffic Planning, Lund.

Integrating fixed and mobile arterial roadway data sources for transportation analysis

Alexander Bigazzi (corresponding author); abigazzi@pdx.edu
Wei Feng; wfeng@pdx.edu
Adam Moore; adam.moore@pdx.edu
Miguel Figliozzi; figliozzi@pdx.edu

Department of Civil and Environmental Engineering
Portland State University
P.O. Box 751-CEE, Portland, OR, 97207
Ph: 503-725-4282

ABSTRACT
Fixed-site and mobile data sources have distinct strengths for representing roadway conditions, and unique insights can be generated by combining the two. Ongoing research on an urban arterial roadway corridor combines distinct data sources for multi-criteria transportation facility performance assessment. The roadway is instrumented with an adaptive traffic signal system collecting signal phase and vehicle count data and mid-block radar detectors recording high-resolution vehicle counts, speeds, and classifications. Transit buses on the corridor record stop-level vehicle position and passenger activity data, and air quality has been measured with deployable roadside monitoring stations and a portable multi-sensor device for travelers. Combinations of these data sources provide new insights about performance of the transit signal priority (TSP) system, relationships between signal operations and roadside air quality, and variations in on-road exposure concentrations.

Keywords: roadway data, performance measures, transit operations, traffic signals, air pollution
1 INTRODUCTION

The most recent U.S. surface transportation act, MAP-21, continues the trend of increasing use of performance measures for analysis of transportation systems. The increasing demands of performance measures along with technological developments have led to an abundance of new data sources available to transportation analysts. Traditional transportation data come from fixed infrastructure elements such as video cameras or embedded inductive loop detectors. Mobile data sources emerged with the gathering of sensor data from fleet vehicles such as transit buses and taxis. The recent proliferation of personal mobile devices expands the opportunities for mobile data collection.

Although there is a profusion of data originating from transportation systems, challenges remain in converting those data into information (Tufte, Bertini, Chee, Fernandez-Moctezuma, & Periasamy, 2010). One underused analytical approach is the integration of fixed and mobile data sources, which each provide a unique view of roadway conditions. Although new analysis techniques might be required, combining fixed and mobile data can provide a fuller and potentially more accurate representation of conditions (Feng, Bigazzi, Kothuri, & Bertini, 2010).

This paper presents findings based on integrated fixed and mobile data sources on an urban arterial roadway. The use of combined data for multi-criteria transportation facility performance assessment is demonstrated and challenges are discussed. Three separate analyses are presented for 1) performance of a transit signal priority system, 2) relationships between signal operations and roadside air quality, and 3) variations in on-road exposure concentrations. The next section describes the data sources on the corridor, and the following section presents performance measurement results.

2 ROADWAY DATA SOURCES

2.1 Description of the study corridor

This research utilizes existing data streams from an instrumented corridor as well as temporary deployments of stationary and mobile traffic, atmospheric, and air quality monitoring systems. The instrumented corridor is SE Powell Boulevard, an urban arterial in Portland, Oregon. Powell Boulevard (also U.S. Highway 26) connects the Portland downtown and the City of Gresham, a suburban city. Powell Boulevard has two lanes of traffic in each direction for most of its length, and a center turn lane or median for left turns in some sections. Sidewalks are present on both sides of the roadway and there are no bicycle lanes. The street route runs east-west and includes the Ross Island Bridge which crosses over the Willamette River (annual average daily traffic at the bridge is over 60,000 vehicles). The study area is highlighted in Figure 1.

Powell Boulevard is regularly congested during peak traffic hours. The morning peak period is in the westbound direction, towards downtown Portland, while the afternoon peak period occurs in the eastbound direction. Powell Boulevard is a key regional commuter facility. This facility is classified as a designated Oregon Department of Transportation (DOT) Region 1 Critical Urban Arterial Corridor. The Ross Island Bridge carries the most traffic volume of any 4-lane facility in the Portland metropolitan area (Figure 2 shows eastbound traffic entering the SE Powell Boulevard corridor from this bridge). Most traffic signals are separated by five or six blocks and in many occasions traffic queues occupy the extension between traffic signals.
Several bus routes run along Powell Boulevard and are affected by congestion, especially the high frequency bus route 9. Route 9 is within the top five TriMet\(^1\) routes in terms of productivity and passenger demand. The peak periods for Route 9 coincide with general vehicle

---

\(^1\) TriMet stands for Tri-county Metropolitan area and is the public transit agency serving the three counties that comprise the Portland metropolitan region with a current population estimated around 2,300,000
traffic peaks and occur in the morning for the westbound direction and in the afternoon for the eastbound direction.

Along Powell Boulevard there is a variety of land uses coupled with intense commercial, educational, recreational, and residential activities. For example, Powell City Park is located between SE 22nd and SE 26th Streets; Cleveland High School (1,500 students) is located between SE 26th and SE 27th Streets and its recreational and sports facilities are located between SE 31st and 33rd Streets. In addition, numerous strip malls, popular restaurants, a brewery, a supermarket, and residential apartment complexes are located along Powell Boulevard. Improving the performance of this arterial is difficult due to the competing needs of different types of users such as pedestrians, transit, and private automobiles as well as balancing mobility and accessibility for a diverse array of activities and land uses along the corridor.

Figure 3. Development zoning along study corridor.

2.2 Fixed traffic data sources

The study corridor was chosen because of a unique set of complementary traffic data technologies. The City of Portland and the Oregon Department of Transportation deployed a state-of-the-art adaptive traffic signal system called SCATS (Sydney Coordinated Adaptive Traffic System-SCATS) between September 2011 and March 2012. SCATS optimizes traffic signals and traffic performance adapting or responding to changes in traffic volumes or conditions along the corridor. The green and red signal durations change throughout the day as SCATS coordinates traffic signals along the entire corridor, enabling the formation of “platoons” of vehicles that move together. Intersection cycle lengths, phase splits, and offsets are optimized on a cycle-by-cycle basis, allowing the entire corridor to adapt to changing traffic trends faster than traditional traffic signal systems. SCATS detectors record vehicle count data for every lane approaching an intersection, as well as all phase timing data.

Digital Wave Radar (DWR) sensors measure arterial traffic volumes, speed, vehicle length, vehicle type, and lane occupancy rate. Two Wavetronix-brand permanent DWR sensors are situated at mid-block locations (between major intersections) along the corridor – one at SE 24th Avenue and one at SE 35th Avenue. DWR sensors quantify vehicle volumes and vehicle types based on length (passenger car versus truck) at the lane level and per direction of travel.

2.3 Transit operations data

The study area includes 22 bus stops in each direction. Route 9 is the primary bus line along the corridor, with additional service from Route 66 and four other routes on cross streets (see Figure 4). Recorded bus stop event data include bus arrival and departure times, passenger activities (boarding and alighting), and vehicle and driver information at each bus stop along the
corridor. Each bus GPS record has a unique ID that can be matched to the bus vehicle data (brand, size, age, mileage, maintenance record, type of UFP filter, actual fuel efficiency, etc.). The bus fleet uses B5 biodiesel (5% biodiesel, 95% petrol diesel).

![Image of transit lines along the study corridor.](Image)

**Figure 4. Transit lines along the study corridor.**

To analyze the transit signal priority system performance, the bus stop event data, SCATS signal phase log data and traffic count data are integrated into one bus stop-to-stop trip database. A bus stop-to-stop trip is a bus trip between two consecutive bus stops that includes one SCATS signal. After data integration, each bus stop-to-stop trip includes attributes include bus stop activities (e.g. arrival time, departure time, schedule delay, passenger boarding/alighting, on-board passengers) at both bus stops, travel time and distance between bus stops and intersections, estimated bus arrival time at signalized intersections, estimated signal delay due to a red signal indication and/or time savings due to a transit signal priority (TSP) phase.

### 2.4 Roadside air quality data

In order to investigate the effects of detailed vehicle activity on near-road pollutant concentrations, high-resolution traffic data from the mid-block DWR detectors were used in concert with similarly high-resolution air quality measurements. Roadside air quality data were collected using portable instruments deployed during peak periods to measure sidewalk-level air pollution exposure concentrations. All instruments were placed on a portable table 2.5m from the roadway (see Figure 5). Data collection took place only in the absence of precipitation.

Data were collected at one second resolution and later aggregated to 10 seconds to match the output of the DWR traffic sensors. PM mass concentrations were measured with the TSI DustTrak (TSI Model 8533). The DustTrak measures three PM size designations: PM$_{1}$, PM$_{2.5}$, and PM$_{10}$, corresponding to PM with aerodynamic diameters below 1, 2.5, and 10 micrometers, respectively. Wind speed and direction were measured using an RM Young Ultrasonic Anemometer (Model 81000), and temperature and relative humidity are measured with a HOBO U12-013 (Onset).
2.5 On-road exposure concentration data

Assessment of air pollution intake requires combined knowledge of environmental, physiological, and travel conditions. Currently there is a lack of tools that allow integrated measurements of these data. In response, a portable, low-cost, multi-sensor device was recently developed and field-tested on the study corridor (as well as on other roadways in the area). The prototype device combines trajectory (location, speed, acceleration), local traffic (passing vehicles), air quality (CO, VOC), meteorology (temperature, humidity), and physiology (heart rate) data for travelers. The device connects wirelessly with a smartphone running a custom application that displays and logs the data, in addition to incorporating information from the smartphone’s GPS receiver and third-party bio-monitors. Post-processing the space/time stamps of these mobile data with the stationary traffic sensors on the corridor allows analysis relating the on-road measurements to traffic conditions. The prototype device is described in the documentation which is available for download along with the microcontroller and application source codes² (Bigazzi, 2013).

3 PERFORMANCE MEASUREMENT

3.1 Transit signal priority (TSP)

Transit signal priority (TSP) is the process of detecting transit vehicles approaching signalized intersections and adjusting the phasing of the signal in real time to reduce the delay experienced by the transit vehicle (Furth & Muller, 2000). The two most common TSP phases are green extension and early green (or red truncation).

Previous studies analyzed the impact of TSP systems on bus travel time savings, on-time performance, headways, and the delay and time savings for other vehicles. Due to the lack of disaggregated TSP phase log data and integrated analysis between signal phase log data with bus automatic vehicle location (AVL) and automated passenger count (APC) data, no study has

² [http://alexbigazzi.com/PortlandAce/](http://alexbigazzi.com/PortlandAce/)
evaluated TSP performance at the TSP phase level, assessing the effectiveness and efficiency of TSP phases for buses that request priority. TSP effectiveness of an intersection measures the percentages of TSP phases that are early, on-time or late to a bus when it arrives at an intersection. For example, it is possible that 20 buses requested priority at an intersection, ten TSP phases were triggered, five, three, and two of them were granted early, late and on-time (beneficial) to the buses, respectively. TSP efficiency of a TSP phase assesses the passenger time savings and vehicle delay per second of TSP phase duration. These disaggregated TSP performance analysis can help the cities and transit agencies to better understand the existing TSP system performance, identify potential problems and improvement opportunities.

Figure 6 shows the average number of bus trips per day that did and did not request TSP from both directions at intersections between 26th Ave. and 72nd Ave. along Powell Blvd. It shows that almost half of the bus trips requested TSP at each intersection. Because there are no bus emitter activation/deactivation records, a bus is determined to have requested priority if the bus actual departure time is more than 30 seconds late than the scheduled departure time.

![Figure 6. Average number of bus trips per day.](image1)

![Figure 7. Average number of TSP phases per day.](image2)

Figure 7 shows the average numbers of green extension phases and early green phases per day. It shows that few TSP phases were granted at the intersections of 26th Ave. and 33rd Ave.
Ave. on Powell Blvd., which indicates a potential TSP setting problem at these two intersections. There are more green extension phases than early green phase. The average durations of green extension and early phases are 7 and 11 seconds, respectively. Figure 6 and Figure 7 show that the average number of TSP requests is much higher than the number of TSP phases at each intersection. Therefore, not all of the TSP requests triggered the granting of a TSP phase.

3.1.1 Data Integration

Integrating the bus AVL/APC data with the SCATS signal log data and vehicle count data is important because it provides the required information for TSP performance analysis. It is also a challenging step because the bus AVL/APC data and SCATS data are collected in different spatial dimensions. Bus AVL/APC data are collected at bus stops while SCATS data are collected at intersections. Bus trajectory information is unknown between bus stops. However, TSP performance analysis at the TSP phase level requires bus arrival time information at the intersections. Therefore, this study developed an algorithm to estimate the probabilistic bus arrival times at an intersection based on: 1) empirical bus travel speed probability distribution; 2) bus departure and arrival time at the upstream and downstream stops of an intersection; and 3) the signal phase start and end times of that intersection. These estimated bus arrival time probability distributions are used to estimate TSP performance measures. Based on this algorithm, we integrated the three data sources into one “bus stop-to-stop trip” database. Each bus stop-to-stop trip is a data record, and each bus stop-to-stop trip is associated with some input information and output attributes. Table 1 lists the input data and output attributes for each bus stop-to-stop trip record. The input data are from the three original data sources, the output variables are estimated based on the algorithm.

Table 1. Bus stop-to-stop trip attributes

<table>
<thead>
<tr>
<th>Input Data</th>
<th>Output Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upstream/downstream distance</td>
<td>Priority request</td>
</tr>
<tr>
<td>Bus departure/arrival time at the upstream</td>
<td>Probabilities of bus arriving time at intersection</td>
</tr>
<tr>
<td>and downstream stops</td>
<td>during green/red/green extension/early green</td>
</tr>
<tr>
<td>Passenger boarding/alighting/load</td>
<td>Bus signal delay</td>
</tr>
<tr>
<td>Signal phase start/end times</td>
<td>Bus time saving due to a TSP phase</td>
</tr>
<tr>
<td>Traffic volumes</td>
<td>Non-bus vehicles time saving/delay due to a TSP phase</td>
</tr>
</tbody>
</table>

In the output variables in Table 1, “priority request” indicates whether a bus requested priority. Given the bus departure and arrival time at the upstream and downstream stops and the signal phase start/end time of an intersection, probabilities of the bus arriving at the intersection during green, red, green extension and early green phases can be estimated. This information can be used to estimate the effectiveness of TSP phases. Based on this information, signal delay for a bus stop-to-stop trip can also be estimated. The expected time saving for a bus due to TSP phase can be estimated, and the expected time saving for other vehicles on the main street and expected delay for other vehicles on the side street can also be estimated.
Figure 8 shows an example of a bus stop-to-stop trip in a time-space diagram. $d_1$ and $d_2$ represent the distance between the upstream bus stop and the intersection stop bar, and the distance between the intersection stop bar and the downstream bus stop; $dt_i$ and $at_i$ are the departure time from the upstream stop and the arrival time at the downstream stop for bus trip $i$; $R_j^s$ and $R_j^e$ are red phase start time and end time in cycle $j$; $GE_j^s$ and $GE_j^e$ are early green phase start time and end time in cycle $j$. Although the true bus trajectory is unknown, based on the bus travel speed probabilistic distribution for each stop-to-stop segment, a feasible bus trajectory boundary can be drawn in the time-space diagram given $dt_i$ and $at_i$. The earliest and latest possible times that bus trip $i$ could have arrived at the intersection given $dt_i$ and $at_i$. Based on the probability distribution of bus travel speeds, the expected signal delay of bus $i$ and the expected time saving due to the green extension phase can be estimated. The equations for estimating these output variables are omitted for simplicity. More details about the TSP analysis can be found in Feng (2014).

3.1.2 TSP Effectiveness

In this study, if a TSP phase is granted in the same cycle when a bus arrives at the intersection and this bus requested priority, this TSP phase is granted in time to the bus that requested priority, and this TSP request triggers a TSP phase granted in time. The timely cycles are defined differently between green extension (GE) and early green (EG) phases so that the TSP phase is close to the middle of a timely cycle. As shown in Figure 9, a timely cycle for a GE phase is the time interval between two consecutive green phase start times, and a timely cycle for an EG phase is the time interval between the middle time of two green phases. In Figure 9 (a) and (b), a bus “d” arrives at the intersection in cycle ① but a TSP phase is granted in cycle ②; therefore, this TSP phase in cycle ② is not triggered by any bus. Bus “a”, “b” or “c” arrives at the intersection in cycle ③ and each could have triggered the TSP phase. This TSP phase in cycle ③ is granted late, on-time and early to bus “a”, “b” and “c”, respectively. However, only when a TSP phase is granted on-time to a bus that requested priority, it is considered effective. Therefore, whenever there is a TSP phase granted, it is possible that: 1) no bus triggered this TSP.
phase within a cycle length; 2) this TSP phase is late to a priority request; 3) it is on-time to a priority request; and 4) it is early to a priority request.

Based on the estimated probabilities that buses arrive at an intersection during the green, red, green extension and early green phases, the percentages that a TSP phase was granted early, on-time, late and in a different cycle to a TSP request are shown in Figure 10 (a) and (b).

Results vary significantly across intersections and by direction. Figure 10 (a) shows that, on average, 64% of the GE phases are late, 28% of them are granted in a different cycle and 5% of them are on-time. This means that 95% of GE phases are not effective and most of them are late. Results clearly indicate a problem with the GE phases. This might be a TSP control strategy problem or a TSP request detection/deactivation problem. Figure 10 (b) shows that, on average, an EG phase has 40% probability of being on-time, 30% probability of being early and 28% probability of being in a different cycle. Therefore, EG phases are much more effective than GE phases.

---

**Figure 9. TSP timeliness and effectiveness illustration.**

Based on the estimated probabilities that buses arrive at an intersection during the green, red, green extension and early green phases, the percentages that a TSP phase was granted early, on-time, late and in a different cycle to a TSP request are shown in Figure 10 (a) and (b).

Results vary significantly across intersections and by direction. Figure 10 (a) shows that, on average, 64% of the GE phases are late, 28% of them are granted in a different cycle and 5% of them are on-time. This means that 95% of GE phases are not effective and most of them are late. Results clearly indicate a problem with the GE phases. This might be a TSP control strategy problem or a TSP request detection/deactivation problem. Figure 10 (b) shows that, on average, an EG phase has 40% probability of being on-time, 30% probability of being early and 28% probability of being in a different cycle. Therefore, EG phases are much more effective than GE phases.
3.1.3 TSP Efficiency

A bus can benefit from a GE phase only if the bus arrives at an intersection during the GE phase; on the other hand, a bus can benefit from an EG phase if the bus arrives at the intersection during regular red time or the EG phase. However, if a bus benefits from a GE phase, the time savings will be the time interval between the arrival time of this bus at the intersection and the end time of the following red phase. If a bus benefits from an EG phase, the maximum time savings will be the EG phase duration. Because red phase duration is longer than EG phase duration in most of the intersections, the time savings for a bus that benefits from a GE phase is usually higher than when it benefits from an EG phase. Therefore, it is important to measure the time savings for buses and onboard passengers per TSP phase. It is also necessary to assess the time savings for non-bus vehicles on the major street and vehicle delays on the minor street due to a TSP phase. Also, because the average GE and EG phase durations are different across intersections, the time savings and delay per second TSP phase will be compared.

For each bus stop-to-stop segment, the average bus passenger time savings per second TSP phase can be estimated by:
Figure 11 (a) and (b) show that the estimated total passenger time savings per second GE phase is much lower than EG phase. Therefore, early green phases are more efficient than green extension phases in most of the intersections. This may be because there are too many GE phases that are not beneficial to any buses.

According to (Smith, Hemily, & Ivanovic, 2005), TSP works better at far-side stops because bus arrival time prediction is more reliable at far-side stops. However, Figure 11 (a) and (b) do not show clear differences between near-side and far-side stops. This finding does not indicate that near-side and far-side stop configurations have no impact on TSP performance, because there are only six near-side stop segments, and five of them may have TSP settings problems.

Figure 11. Estimated total passenger time savings per second TSP phase.
granted, the total time savings (TTS) for non-bus vehicles on the major street and the total delay (TD) for vehicles on the side street can be estimated by the following equations:

\[
TTS = \frac{q_1 \cdot q_2}{2(q_2 - q_1)} (2 \cdot Red \cdot TSP - TSP^2) \tag{2}
\]

\[
TD = \frac{q_1 \cdot q_2}{2(q_2 - q_1)} (2 \cdot Red \cdot TSP + TSP^2) \tag{3}
\]

Assuming all non-bus vehicles are single occupancy vehicles, results are shown in Figure 12. Results show that the total time savings and delays for non-bus vehicles per second GE phase and per second EG phase are very similar (less than 2 seconds difference), which means the nonlinear effect of TSP phase duration on non-bus vehicles time savings and delays is negligible.

**Figure 12. Total passenger time savings and vehicle delays per second TSP phase.**

*Integrating fixed and mobile roadway data, Bigazzi et al.*
For each second EG phase, the bus passenger time savings is slightly less than the total vehicle delay on the side street for intersections west of 52nd Ave., but the sum of the bus passenger time savings and the total vehicle time savings on the major street is much higher than the side street vehicle delay at all intersections. For each second GE phase, the sum of bus passenger time savings and non-bus vehicle time savings on the major street is almost equal to the vehicle delay on the side street.

3.2 Signal operations and roadside air quality

The relationship between traffic operations and air quality was studied using aforementioned air quality, meteorological, and traffic measurement devices. Specifically, fine particulate matter (PM$_{2.5}$) was measured at a mid-block location between two traffic signals during a morning peak period, directly in line with a DWR sensor measuring vehicle volume, speed, and lane occupancy. More details about this analysis can be found in Moore et al. (2014).

Data were analyzed at 10-second intervals, a much finer resolution than comparable roadside air quality study designs. This resolution allowed special attention to be paid to changes in traffic conditions, including fleet mix, queuing, and vehicle platooning. Significant correlations were observed between vehicle platoons and increases in PM$_{2.5}$ concentrations.

During the data collection period, substantial congestion in the morning commute direction (westbound) resulted in a breakdown in traffic flow. Average westbound speeds were about 30mph at the start of the study period (7:00am), but quickly dropped to below 10mph and did not improve by the end of the study period (9:00am). The availability of high resolution traffic data from the DWR sensor allowed the investigation of traffic state-pollutant concentration analysis. Two types of traffic states were explored: the onset of congestion and the periodic arrival of vehicle platoons from upstream signals.

3.2.1 Congestion Identification

Traffic volume data indicate the number of vehicles present near the air quality monitoring site, but they do not necessarily indicate the traffic state. That is, whether a roadway is congested or not. A method derived from Bertini (2003) was employed to empirically identify traffic states in which two states are identified: congested and uncongested. Cumulative speeds, $N(t)$, were plotted against time (top row in Figure 13). The curve’s slope at time $t$ represents the vehicle speed at that time. A rescaled cumulative (oblique) speed curve (bottom row in Figure 13) amplifies changes in speed to emphasize the moment of speed reduction. The oblique speed curve was created by reducing $N(t)$ from $v_0 t$, where $v_0$ is an oblique scaling rate.
Figure 13  Rescaled cumulative (oblique) speed curve construction using cumulative speeds, $N(t)$, and oblique cumulative speeds, $N(t) - v_0t$. Shaded area indicates active queue.

A local maximum on the oblique speed curve indicates a time at which a speed reduction occurred, and a local minimum indicates a time at which a speed increase occurred. The two conditions are, in this context, referred to as queuing activation and deactivation points, respectively. The shaded portion of the oblique speed curve indicates an active queue. The onset of congestion was identified at 7:38am. From that point, the queue was primarily active, indicating congestion. No congestion occurred in the eastbound direction.

Median PM$_{2.5}$ concentrations were compared before and after the onset of congestion (see Figure 14). Boxplots of the concentrations indicated lower median concentrations for active queuing periods prior to congestion than for inactive periods. In contrast, higher median concentrations were observed for active queuing periods after the onset of congestion than for inactive periods.
Active queuing periods after the onset of congestion may differ from those prior to congestion, helping to explain the discrepancy in median PM$_{2.5}$ concentrations. Prior to the onset of congestion, active queuing periods are characterized by brief decreases in speed, though for time durations too short to bring traffic to congested conditions. Short queuing periods outside of congestion, then, likely lead to traffic conditions with lower accelerations, which may result in lower emissions rates and lower PM$_{2.5}$ concentrations.

3.2.2 Vehicle Platoon Identification

The cyclic nature of vehicle presence in the study due to the upstream traffic signals was detectable in the traffic volume data using autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. ACF and PACF plots illustrate the similarities between observations as a function of the time lags between the observations (see Figure 15). Both of these functions are commonly used as a method for determining repeating patterns.
Integrating fixed and mobile roadway data, Bigazzi et al.

Figure 15  Autocorrelation of vehicle volumes showing vehicle platooning with upstream signal cycles (1 lag = 10 seconds)

The westbound upstream signal had a median cycle length of 123 seconds, and the eastbound signal had a median cycle length of 125 seconds. Thus, vehicle platoons would be expected to arrive every 123 and 125 seconds from their respective directions. These cycles are evident in the vehicle volume ACFs and PACFs in Figure 15. One lag in the figure equals 10 seconds, and the eastbound response has a clear spike at 13 lags, or 130 seconds, due to traffic platooning. The westbound direction had a slightly more dampened response, likely due to the substantial congestion, which mitigated any upstream cycle effect due to the constant vehicle presence at the sensor location.

To investigate the effect of vehicle platooning on roadside PM$_{2.5}$ concentrations, a cross-correlation function (CCF) was made for both directions of travel. The CCF in Figure 16 illustrates correlations between PM$_{2.5}$ concentrations at time $t$ and traffic volumes at time $t + h$, for $h=0, \pm 1, \pm 2, \pm 3, \ldots$ Negative values for $h$ indicate a correlation between volumes at a time $h$ units before $t$ and PM$_{2.5}$ concentrations at time $t$. The dashed lines indicate the statistical significance level. No westbound correlations were significant; westbound vehicle platooning did not have a significant effect on PM$_{2.5}$ concentrations. In the eastbound direction of travel, PM$_{2.5}$ concentrations were significantly positively correlated (+10.4%) with vehicles passing at 12 lags, or 120 seconds. This lag time roughly matched the upstream cycle length, and the eastbound vehicle ACF and PACF. Eastbound vehicle platooning, then, significantly positively correlated with PM$_{2.5}$ concentrations. It is likely that platooning conditions were easier to
identify in the eastbound direction due to the uncongested conditions and clear vehicle arrival times for the duration of the study period.

![Cross-correlation of vehicle volumes with PM$_{2.5}$ concentrations (1 lag = 10 seconds)](image)

**Figure 16** Cross-correlation of vehicle volumes with PM$_{2.5}$ concentrations (1 lag = 10 seconds)

### 3.3 On-road exposure risk and traffic conditions

Travelers’ health risk from air pollution is one dimension of a roadway’s environmental performance. Although most research focuses on exposure concentrations, ventilation is also an important determinant of pollutant inhalation and uptake – particularly for active travelers such as pedestrians and bicyclists (Bigazzi & Figliozzi, 2014). In order to connect a roadway’s functional and operational condition with pollutant inhalation risks, three dimensions of data are needed:

1. Travel and traffic conditions (location, facilities)
2. Environmental conditions (pollutant concentrations, weather)
3. Physiological conditions (traveler’s ventilation rate)

Figure 17 is an example of the information that can be gained from integrating high-resolution environmental and physiological measurements. Figure 17 shows second-by-second data from a 7.9 km (4.9 mi) bicycle trip along the study corridor starting from an inner commercial/industrial area (from right to left) on July 11, 2013, with the measured exposure shown as the pin height and the bicyclist’s breathing rate shown as the pin color (with lower to higher rates shown as red-yellow-orange-green). Exposure “hot-spot” areas with jointly higher ventilation and concentration values can be identified by tall green pins. The sample trip in Figure 17 started with high concentrations but relatively low ventilation in the commercial/industrial area. The largest exposure “hot-spot” occurred after crossing SE 39$^{th}$ Ave (marked on the figure with a red box). This location has high traffic volumes and an upward grade, where we would expect both greater vehicle emissions and greater bicyclist exertion.
In order to examine the effect of traffic on exposure risks, the on-road data were mapped to the roadway network and synchronized with the traffic data. GPS data points were mapped by proximity to a link-based GIS layer from the City of Portland that included estimated ADT for each link based on interpolated traffic count data. Figure 18 shows correlation coefficients between ADT and 10-second aggregated exposure and physiology data from 12.5 hours on a variety of mixed-traffic roadways in the area (not only the Powell Blvd. study corridor, which has roughly consistent ADT along its length). All data were collected between April and September, 2013 near the morning peak period (7-10 am). On-road exposure is shown for carbon monoxide (CO) and volatile organic compounds (VOC) sensors; on-road physiology is shown for heart rate and breathing rate. Correlation coefficients are indicated by the color shading of the cell, and significant coefficients based on two-tailed t-tests ($p < 0.05$) are also printed in text.

The two exposure variables are significantly positively correlated with each other, as are the two physiological variables, as expected. Exposure and physiology are both positively correlated with ADT, though the correlation with physiology is stronger. Bicyclist self-determined speed was found to be higher on larger roads, which is consistent with the physiology/ADT relationship.

Concurrent traffic conditions for the on-road data were determined by matching time-stamps with the DWR data set described above. Figure 19 shows correlation coefficients among three DWR traffic variables (volume, speed, and density) and the four on-road data variables from the previous figure. A subset of only on-road data from the study corridor is included (1.5 hours). Correlation coefficients are indicated by the color shading of the cell, and significant coefficients ($p < 0.05$) are also printed in text.

Amongst the traffic variables, density was significantly correlated with volume and (negatively) with speed. The insignificance of the volume-speed correlation could be due to the non-linear shape predicted from traffic flow theory (May, 1989). Traffic speed was negatively correlated with exposure concentrations; concentrations were higher during congestion when speeds were low. Traffic volume and density were only significantly correlated with CO, not VOC. Traffic volume and density were positively correlated with physiology, though the traffic speed/physiology correlation was not significant. Overall, both high-volume facilities and congested periods (with low traffic speeds and high traffic density) were associated with higher pollution risks due to exposure concentration and ventilation.
Figure 18. Correlation coefficients for 10-second aggregated data from all mixed-traffic roadways (significant relationships at $p<0.05$ are marked with text; $N=4,500$)
4 CONCLUSIONS

This paper presents three analyses with unique findings that arise from combining fixed and mobile transportation data sources. Assessment of the performance of a transit signal priority system required a combination of traffic signal timing data and on-bus location data to determine the signal status when buses arrived at intersections. Vehicle volume data were also used to quantify the time trade-offs for bus and auto passengers on the major arterial road and cross roads. A study of roadside air quality combined fixed-site traffic data with data from a deployable air monitoring station to quantify the high-resolution impact of traffic characteristics on near-road air quality. The third analysis combined on-road mobile air quality and physiology data with stationary traffic data to reveal correlations between on-road exposure risk and traffic characteristics across both space (ADT on different links) and time (varying traffic levels on one facility).

Fixed and mobile data integration requires synchronization and aggregation in both space and time dimensions. The data processing for these analyses was \textit{ad hoc} and manual. In order to realize the potential of emerging transportation data sources, the combination of high-resolution fixed and mobile data types must become a less resource-intensive activity. One possibility for facilitating these data integrations is increased flexibility and use of archived data user services (ADUS) such as PORTAL in Portland, Oregon (Tuft et al., 2010). ADUS architectures could...
aid analysts by providing enhanced synchronization, aggregation, and integration functions for a wider range of data sources.

5 REFERENCES
Inside the Triangle:
Does Database Selection Alter our Understanding of Urban Industrial Systems?

Mary Donegan
Doctoral Candidate
Department of City and Regional Planning
University of North Carolina at Chapel Hill

Workshop on Big Data and Urban Informatics
August 11-12, 2014
University of Illinois at Chicago

Preliminary research
Please contact author prior to citation for an updated version

Support for this research was provided by the National Science Foundation and the Study for the Center of the American South at the University of North Carolina at Chapel Hill. While the research has benefited from the work and input of many, special thanks to Dr. Nichola Lowe, Dr. Maryann Feldman, and Cara Wittekind.
Abstract: Big Data promises to deepen our understanding of urban industrial systems by providing insight into industrial dynamics at far finer geographic scales than previously possible. Yet, our ability to achieve the promises that fine-grained data hold have been stymied by: (1) our long-standing inability to access fine-grained, government-produced data, and (2) data quality concerns with the growing catalog of alternative data sources. Previous research suggests there are significant inaccuracies in alternative data sources, and that inaccuracies may bias study conclusions. More nuanced examinations of alternative industrial data suggest that we have paid insufficient attention to purposeful differences, which are driven by divergent database design practices utilized by government agencies and private companies. Conflating inaccuracies with purposeful differences decreases our understanding of true inaccuracies and the appropriateness of each database for specific research questions.

This research provides exploratory research into the observed differences between entrepreneurial records in government and alternative data sources. The research compares data from the Quarterly Census of Employment and Wages (QCEW) against a leading private alternative, the National Establishment Time Series database (NETS). The research asks the questions: To what extent do entrepreneurial records in the QCEW and the NETS diverge? How much is driven by purposeful differences, as opposed to inaccuracies? What methods can be used to assess differences between entrepreneurial records in the databases? How does a reliance on one database over another alter our understanding of a region's entrepreneurial economy? The research analyzes the entrepreneurial economy in North Carolina's Research Triangle's biomedical industry, paying special attention to differences in establishment inclusions, the timing and trend of establishment births, initial employment levels at these establishments, and spatial patterns of entrepreneurial establishments. The research finds that while inaccuracies in the NETS exist, purposeful differences drive differences between the two databases.

Key Words: secondary data; data quality; entrepreneurship; industrial clustering
Introduction

Big Data promises to deepen our understanding of urban industrial systems by providing insight into industrial dynamics at far finer geographic scales than previously possible. By analyzing a region’s fine-grained industrial data—such as information on firm births, employment levels, and locations—we may be able to both better identify spatially heterogeneous industrial trends within a region and test neighborhood-level impacts of economic development policies. Yet, two major issues limit our ability to achieve the promises that fine-grained data hold: (1) our collective and long-standing inability to access the fine-grained, government-produced industrial data that is considered the research "gold standard," and (2) data quality concerns with the growing catalog of readily-available, alternative data sources developed by private companies. Researchers stymied by their inability to access government sources have increasingly relied on these alternative sources, even as a lingering inability to fully grasp the scale and scope of alternative data inaccuracies undermines confidence in study conclusions and resulting policy recommendations.

Previous research suggests that there are significant inaccuracies in alternative data sources, and that these inaccuracies may bias study conclusions and policy recommendations (see, for example, Fleischhacker et al., 2012; Liese et al., 2010; Ma et al., 2013). Yet, more nuanced examinations of alternative industrial data suggest that while inaccuracies exist, we have also paid insufficient attention to purposeful differences between government and private databases (Kunkle, 2011; Neumark, Zhang, & Wall, 2005). These purposeful differences are distinct from inaccuracies. They are driven not by insufficient data collection efforts employed by private companies, but rather by divergent database design practices utilized by government agencies and private companies. This distinction is a critical one. Conflating inaccuracies with purposeful differences decreases our understanding of both true inaccuracies in alternative databases and the appropriateness of each database for specific research questions. Without separating these twin drivers of differences between government and alternative databases, we remain unable to fully assess both the appropriateness of alternative data sources for our research and the validity of study conclusions built on alternative data analysis.

The goal of this research is to provide exploratory research into the observed differences between government and alternative data sources. Using start-up entrepreneurial firms in the biomedical industry as a case study, the research compares data from a well-known government source for industrial and employment data (the Quarterly Census of Employment and Wages [QCEW]) and a leading private alternative (the National Establishment Time Series database [NETS]). Using these databases, the research asks the questions: To what extent do entrepreneurial records in the QCEW and the NETS diverge? How much of this difference is driven by purposeful differences between the two databases, as opposed to inaccuracies in the NETS? What methods can be used to assess differences between entrepreneurial records in the QCEW and the NETS? And finally, how does a reliance on one database over another alter our understanding of a region’s entrepreneurial economy?

To answer these questions, the research analyzes the entrepreneurial economy in North Carolina’s Research Triangle’s biomedical industry. Entrepreneurial firms are drawn from a unique database developed at the University of North Carolina at Chapel Hill, which includes extensive information on each entrepreneurial firm (Feldman & Lowe, Forthcoming). The research examines establishment-level records for these entrepreneurial firms in the QCEW and
the NETS, with special attention paid to differences in establishment inclusions, the timing and
trend of establishment births, initial employment levels at these establishments, and spatial
patterns of entrepreneurial establishments in the region. Consistent with the literature, these
methods ultimately find that the NETS captures substantially more establishments than the
QCEW, with earlier founding years, higher levels of employment per company, and greater
spatial diffusion across the region. However, the research also shows that removing known
entrepreneurial employment reduces the average difference between NETS and QCEW
employment records by over 60 percent, and that the spatial concentration found in the QCEW is
largely due to its exclusion of establishments at residential addresses. In total, the research finds
that while inaccuracies in the NETS exist, purposeful differences drive differences between the
two databases.

The remainder of the paper is laid out as follows. The next section outlines key theoretical and
practical differences between the QCEW and the NETS, particularly as they relate to
entrepreneurial firms, and summarizes key pieces in the literature that assess the accuracy of the
NETS. The paper then introduces the biomedical industry in North Carolina's Research Triangle,
before providing an overview of the methodology used to assess differences between the
databases. The paper concludes with a discussion of the results, future research steps, and
preliminary lessons that can be drawn from the research.

The QCEW versus the NETS: A Comparison

Background Information on the QCEW and the NETS: The Quarterly Census of Employment
and Wages (QCEW, formerly known as the ES-202) is the product of a partnership between the
federal government's Bureau of Labor Statistics (BLS) and each State Employment Security
Agency (SESA; specific agency names vary by state and often change over time). Each state
uses the QCEW as a record of jobs each company has that are eligible for unemployment
insurance (UI). Each SESA collects data on establishments' addresses, monthly employment,
NAICS codes, and wages. SESAs send their collected information to the BLS, which in turn
releases the data to the public.

However, the data the BLS makes publically available is not equivalent to the data it receives
from SESAs. To protect confidentiality, the BLS will not release an establishment's individual
information. In most cases, the BLS protects confidentiality by aggregating QCEW data to
county, metropolitan, state, and national geographic areas. Within these areas, users can examine
data by industry using NAICS codes. Aggregation conceals gross establishment additions and
closures, establishment-level employment changes, and physical relocations, all of which are
critical to understanding industrial systems. Furthermore, within a specified NAICS code, the
BLS occasionally suppresses employment and average wage data to protect establishment
confidentiality—in these cases, the agency only releases the number of establishments. In
practice, suppression occurs when there are too few establishments in a given geography's
NAICS code or when too few establishments dominate the category. Together, geographic

1 For example, this study's North Carolina data was collected by the Employment Security Commission; following a
merger, the data is currently housed at the state's Department of Commerce.
2 Collected addresses vary by state. North Carolina, for example, collects each establishment's mailing address, physical
(location) address, and an address for unemployment claims.
3 The North American Industrial Classification System (NAICS) classifies establishments by their economic activity, and
is frequently used by agencies and researchers alike in analyses and assessments of a geographic area's economy.
aggregation and data suppression preclude research on a host of topics—including entrepreneurship, relationships between firm size and employment growth, firm relocation patterns, and firm survival.  

Both researchers and practitioners alike have found it difficult to access the QCEW's disaggregated data (i.e. the establishment-level data before geographic aggregation and suppression). Historically, researchers either 1) used aggregate data, potentially diluting research findings or requiring alternate research questions, or 2) embarked on a lengthy and often unsuccessful application process for access to disaggregated QCEW data. More recently, researchers have embraced a third path: private databases. While researchers have used alternative industrial data sources in the past, the number and diversity of researchers using alternative data sources has increased thanks to the proliferation of these databases, improved database structure, increased accessibility, and perceived improvements in data quality (see Kunkle, 2011 and Neumark et al., 2005 for discussions).

The National Establishment Time Series (NETS) is one of the leading data alternatives to the QCEW. The NETS is a private dataset, built by Walls & Associates using Dun and Bradstreet data (D&B). Like the QCEW, D&B includes information on establishment location and employment levels; unlike the QCEW, D&B does not include wage information. Further unlike the QCEW, which receives its information directly from firms, D&B culls their information from court filings, press releases, newspaper articles, and state filings, and further supplements with over 100 million phone calls to companies (Neumark et al., 2005). Dun and Bradstreet cannot compel companies to answer questions. Yet, by not promising confidentiality, D&B can release data in its raw form without any geographic aggregation or suppression. This serves the purposes of its customers; while QCEW exists to track establishments that must pay UI, D&B sells its information to businesses for marketing and decision-making purposes.

Walls & Associates is a unique user of the D&B data, taking yearly D&B snapshots and creating a longitudinal database from them. Walls & Associates adds information on changes in corporate parentage, and keeps careful track of establishment relocations. Unlike the firewalled QCEW, researchers may be able to access D&B data and the NETS through universities or by purchasing the data.

Over the past decade, private databases have been used in dozens of studies across a range of disciplines. The NETS, for example, has been used to assess the impact of living wages laws on employment outcomes (Lester, 2011), estimate differential impacts of non-profit and for-profit entrepreneurial incubators on firm success (Amezcua, 2010), test the differences of employment growth rates amongst small and large establishments (Neumark, Wall, & Zhang, 2011), and evaluate the role state-level development incentives play in employment growth (Lester, Lowe, & Freyer, 2012). Its source D&B data has been used extensively in assessments of food deserts by researchers in the public health field (see, for example, (Fleischhacker et al., 2012; Liese et al., 2010, 2013; Ma et al., 2013; and Powell et al., 2011).

---

4 County Business Pattern (CBP) is a popular alternative to the QCEW, as the data aggregates to the geographically-smaller zip-code level. However, CBP data has an additional year's delay, pulls from a smaller percentage of establishments, lacks wage data, and does not solve the problem of concealing gross establishment and establishment changes and relocations. Finally, as a database that is aggregated to a smaller geographic level and contains fewer establishments relative to the QCEW, it is prone to intense data suppression (though there have been efforts to fill in the suppressed data; see Isserman & Westervelt, 2006 for work on unravelling CBP suppression).
Even as an increasingly diverse field of researchers and practitioners turn to private establishment-based data sources, questions have arisen about the appropriateness of using alternative data sources. For example, early studies found that D&B identified very low levels of new firm formations (Aldrich, Kalleberg, Marsden, & Cassell, 1989; Birley, 1984; Neumark et al., 2005). More recently, the public health field has suggested that D&B has far lower levels of accuracy when tracking food outlets when compared to physical censuses (see Fleischhacker et al., 2012 for a thorough review and analysis). In short, while many of these authors recognize the need for alternative data, many stress that its accuracy must be assessed.

Yet, a careful examination of the inclusion criteria employed by NETS and the QCEW—that is, the rules each database uses to determine whether or not to include workers and establishments—suggests that while inaccuracies are a problem, they may not be the sole or even leading source of differences between these databases. Overlooking the purposeful differences that result from divergent criteria may inflate the specter of private data inaccuracies and diminish concerns over real differences inherent to the databases.

Inclusion Criteria in the QCEW and the NETS: As a record of jobs employers must pay unemployment insurance on, the QCEW includes data only on workers covered by either state unemployment insurance laws or by the federal Unemployment Compensation for Federal Employees program (UI, collectively). To be included in the QCEW, these covered workers must be either working or receiving payment for work performed in the pay period that includes the twelfth of the month.5 The BLS estimates that 98 percent of paid workers are included in the QCEW (BLS handbook of methods, 1997).

However, two key portions of the labor force are not eligible for UI, and are therefore purposely excluded from the QCEW: unpaid workers and the unincorporated self-employed. Unpaid workers, by nature of not receiving a paycheck, are not eligible for unemployment pay. The BLS classifies self-employment into two categories: the incorporated self-employed, who work at incorporated firms, and the unincorporated self-employed, who work at firms that have not been incorporated. The incorporated self-employed must also pay unemployment insurance on their own wage labor to be eligible for UI; the unincorporated self-employed do not pay unemployment insurance on their own wage labor and are not eligible for UI. In 2009, seven percent of the US workforce fell into the category of unincorporated self-employed (Hipple, 2010).6

In contrast to the QCEW, D&B asks companies how many people work at an establishment, with no respect to time frame, employment status, or payment arrangement (Neumark et al., 2011). As a result, the D&B—and by extension, the NETS—has a much broader definition of employment, making a direct comparison between the two difficult.

These purposeful differences in included workers extend beyond employment counts, as Figure 1 illustrates. The QCEW captures all establishments and workers in incorporated firms comprised entirely of covered workers (box 1). The database also captures all establishments that employ

5 This criteria includes workers on paid sick leave and vacation but not workers on furlough or unpaid absence.
6 It is possible that estimates of unincorporated self-employment are undercounted. Official statistics capture a worker's primary job, not his or her secondary job. In 2009, 1.9 million workers held a second self-employed job (Hipple, 2010). Still, the number of unincorporated self-employed workers has been falling for the past forty years, due to both shifts out of agriculture and an increasing ease to incorporate a business.
any covered workers, but only captures the covered portion of the workers at these establishments (box 2). Critically, the QCEW does not include any establishments without covered workers (box 3).

**Figure 1: Establishments and Workers Included in the QCEW and the NETS**

In contrast, the NETS includes any establishment with a DUNS number (the unique number assigned to any establishment in the D&B database). This inclusion strategy provides for a far broader conceptualization of both a worker and an establishment, as it includes establishments that would be considered non-employer establishments by the QCEW (i.e. establishments with no workers). In theory, unlike the QCEW, the NETS should capture all establishments and workers for all three types of firms.

The importance of these distinctions between workers covered in each of the two databases cannot be overstated. Research comparing the NETS to the QCEW needs to separate (a) what the NETS *should* include and actually includes (i.e. all establishments and all workers vs. the records in the database) from (b) differences between the NETS and the QCEW that are the result of the QCEW's purposeful exclusion of them.

**Accuracy of D&B and the NETS:** A number of studies have assessed the accuracy of data contained in D&B and, by extension, the NETS. Early criticisms of D&B data focused on the low "match rate," or the low percentage of matching records researchers found for externally-defined companies. These studies relied on lists drawn from phone books or ES-202 data (the predecessor name for the QCEW). Birley (1984), for example, found D&B captured just 4 percent of new businesses between 1977 and 1982 in St. Joseph County, Missouri that the ES-202 identified. In their study of new firms in Durham County, North Carolina, Aldrich et al. (1989) found D&B files missed 58 percent of new firms in the ES-202 files and 90 percent found using an enumeration/phone book method. However, they suggest that the ES-202 performed no better, missing 56 percent of D&B firms and 86 percent of firms from enumeration (though the ES-202 data was at an unfair advantage, as it was a year out of date). They also found that D&B
tracked well with data on founding dates and initial employment levels gathered from in-person interviews, a physical census, and the phone book.

Yet, as Neumark et al. (2005) note, both studies were conducted before a critical 1992 infusion of new data into D&B. Using only post-1992 data and robust methods to ensure that reference lists were as accurate as possible, Neumark et al. (2005) found a 95 percent match rate in the NETS for verified new business from a reference database of biotech companies, with 75 percent of the NETS records having the correct founding year and 92 percent being within two years. They also found that while the NETS had significantly higher levels of employment when compared to corresponding industrial and geographic slices in the QCEW, that the majority of employment excess could be removed by subtracting one employee per company (i.e. a proxy for non-wage entrepreneurial labor). Ultimately, many researchers have concluded that the NETS data is a viable alternative to the QCEW and related government databases for studies of long-term trends, and that for entrepreneurship studies in particular, it may be superior (Kolko & Neumark, 2007; Kunkle, 2011; Neumark et al., 2011, 2005).

Recent research in the public health field has been less sanguine towards alternative data sources. These studies of food deserts rely on extensive "ground-truthing," or going to great lengths to ensure the reliability of reference lists of firms by undertaking an extensive physical census of the study area. In their carefully identified list of food establishments in American Indian communities in North Carolina, Fleischhacker et al. (2012) found that D&B had consistently low scores on various measures of accuracy, particularly when compared against Reference USA (a similar alternative database). In their assessment of how various databases predicted low food access areas, (Ma et al., 2013) found that D&B identified fewer of these areas than both InfoUSA (another alternative database) and a reference database. Liese et al., (2013) found that D&B both undercounted and overcounted different types of food establishments to varying degrees, and Liese et al., (2010) found that the D&B only correctly located food establishments within 100meters of the correct location approximately 29 percent of the time. Consensus among these studies is that secondary data sources are not appropriate for identifying geographically and firm-count sensitive sensitive food deserts, and that ground-truthing—or taking a physical census of establishments in a study area—remains the preferred method (see also Powell et al., 2011).

Unfortunately, the robust studies conducted in the public health field contain a series of potential shortcomings when applied to entrepreneurial research. First, methods the public health studies employ to both identify and ground-truth firms limit their assessments to particular industrial categories. Food establishments rely on physical visibility for their long-term viability, making a physical census a valid way to assess their existence and location. Yet, companies not aimed at the general public do not need to have a solid physical presence. Their long-term success may be best built through industry associations and social networking. Furthermore, many entrepreneurial companies are housed in incubators or home offices, often without physical markers of their presence. While there clearly needs to be a reliable reference list of entrepreneurial firms, how to best do this is not as clear-cut as with food establishments.

Second, these studies do not compare either D&B or NETS data against disaggregated versions of the "gold standard" secondary databases like the QCEW. In public health studies, this is likely due to the recognition that physical ground-truthing is the best reference for food establishments. However, it should be noted that both Fleischhacker et al. (2012) and Liese et al. (2010) found that government-produced secondary sources from North Carolina's Department of Agriculture
and Consumer Services and South Carolina's Department of Health and Environmental Control, respectively, performed poorly, too. Neumark et al. (2005) were only able to compare the NETS against aggregated QCEW; (Birley, 1984) and Aldrich et al. (1989) are notable exceptions, though their reliance on pre-1992 data casts doubt on the current applicability of their results.

Third, as the goal of the bulk of these recent studies has been to assess the accuracy of the existence and location of food establishments, they have not focused on employment. Finally, while almost all studies on food deserts measured distances between ground-truthed and D&B data and suggested that alternate data sources would disrupt food desert conclusions, to my knowledge (Ma et al., 2013) are the only authors to carry out an analysis of how various databases alter the understanding of a neighborhood or region.

The Biomedical Industry in North Carolina's Research Triangle: A Case Study

The Industry and its Entrepreneurial Firms: North Carolina's biomedical industry is long-established in the Research Triangle region. The literature has identified the industry's cluster as one of the largest in the country, through both quantitative cluster identification and industry case study methods (e.g., Feldman & Lowe, 2011; Feser, Sweeney, & Renski, 2005; Goldstein, Feser, Freyer, Gordon, & Weinberg, 2008; Markusen, 1996). Researchers have traditionally characterized the industry's local development as the result of a state-led effort that resulted in the founding of the Research Triangle Park and the eventual recruitment of satellite branch plants (see, for example, Luger & Goldstein, 1991; Markusen, 1996). More recent scholarship has reframed the development of the region in a number of ways, including an increased recognition of the region's entrepreneurial tradition (see, for example Feldman & Lowe, Forthcoming; McCorkle, 2012). Entrepreneurial firms in the region have benefited from innovations, formal technology licensing, and labor sourcing from branch plants of multi-national corporations (e.g., GlaxoSmithKline, Pfizer), research findings and labor sourcing from the region's three leading research universities (the University of North Carolina at Chapel Hill [UNC], Duke University, and North Carolina State University), and the guidance and leadership of various non-profit and quasi-public organizations and incubators.

Methodology

The research's overarching approach is to re-create the biomedical industry's entrepreneurial environment in each of the two databases, and then compare the regional pictures each creates. As such, the research starts with an externally defined, fully vetted reference list of the region's entrepreneurial firms in the biomedical industry. Relying on an external reference list follows the approach taken by many of the leading studies (Aldrich et al., 1989; Fleischhacker et al., 2012; Liese et al., 2010; Neumark et al., 2005). Fully triangulating data avoids what these studies have identified—namely, that the sources used to test private data sources, whether phone books or government-provided lists, are often themselves inaccurate, leading to problematic assessments of alternative data sources. It is also more practical for entrepreneurial studies, as secondary databases perform to varying success rates at identifying firms that are both in the biomedical industry and entrepreneurial.

After identifying these firms exogenously from the QCEW and the NETS, the research searches for records of these companies in each of the two databases, compares their records to each other
in an attempt to separate accidental and purposeful differences, and examines differences in temporal trends and spatial concentration. Each of these is addressed in turn below.

Creating a Reference List of Start-up Entrepreneurial Firms in the Triangle's Biomedical Industry: Identifying and tracking the genesis of the region's entrepreneurial firms is the key focus of a multi-year research project at UNC. Led by Dr. Maryann Feldman and Dr. Nichola Lowe, the Circling the Triangle project began with the transfer of a database of 1,800 entrepreneurial firms originally started by the late UNC Chemistry Professor and administrator Dr. William (Bill) Little. Since that transfer, the project has engaged in an ongoing effort to add new firms in innovative industries and expand our knowledge of firms in the database. For each firm, the research team compiles information on: firm entrepreneurs, including their work and educational histories; mergers, acquisitions, and other firm events; patents, trademarks, and licensing agreements; basic firm-specific information, such as location and date founded; and physical relocations, among others.

The result is the Bill Little Database (BLD), which currently includes over 4,000 firms in a variety of innovative industries (e.g., software, clean energy, bio-pharmaceuticals, medical devices; see Feldman & Lowe, Forthcoming). As outlined in Feldman & Lowe (Forthcoming), firm information must be identified from at least three sources; examples of these sources include LinkedIn, in-person interviews, Secretary of State filings, patent filings, incubator tenant lists, technology transfer office filings, and membership or event attendance lists from local entrepreneurial and industry organizations.

This paper's research relies on start-up entrepreneurial BLD firms that fall into the biomedical industry. Most of these firms are engaged in traditional life science or pharmaceutical activities, including early to mid-stage drug development, drug commercialization, contract research, medical device development, and drug and device manufacturing. However, this industry also includes companies that engage in less traditional but supportive activities, including drug development database management software and specialized law and consulting. The result is a more holistic approach than one based on using industrial codes (e.g., NAICS) to define an industry, and an accurate representation of the region's entrepreneurial environment of biomedical start-up firms.

Finally, this research focuses on entrepreneurial firms started from scratch by one or more entrepreneurs, but excludes new, non-entrepreneurial firms. Firms must also have had their first location in the core Research Triangle counties of Durham, Orange, and Wake (see Figure 2). While there are competing ideas about the timing of a firm's emergence (see Aldrich, Kalleberg, Marsden, & Cassell, 1989 and Birley, 1984 for a discussion), this research relies on the earliest-known date that a firm existed, regardless of organizational structure. A firm does not need to have been incorporated, nor does it have to have conducted any formal business transactions. Rather, firms are included for their potential to either conduct business or influence innovations in the regional economy. Firms do not need to have remained entrepreneurial firms after their founding, and in fact many firms in the sample have undergone firms events since their founding (e.g., mergers, acquisitions, IPOs, relocation from the region or state).


dd

Firms founded using technology licensed from an existing company or laboratory are included, but firms spun out of existing firms or that are joint ventures between two existing firms are excluded.

---

\( F_{\text{founded}} \)
Figure 2: The Research Triangle region of North Carolina, as defined by Orange, Durham, and Wake Counties

Finding Full Matches in the QCEW and the NETS: Matching companies from a reference list to their corresponding records in a secondary sources presents significant challenges. At best, minor name changes—such as shifts from LLC to Incorporated, or dropping punctuation from a name—make matching time-consuming. At worst, major historical name changes that are the result of company events (e.g., mergers, acquisitions) or a recognition on the part of the entrepreneurs that the initial name is confusing, a poor representation of the firm's activities, or too common can make matches from a reference list to a secondary source impossible. However, many studies, while thoroughly documenting their statistical methods, have not given similar attention to their matching methods. This is problematic, especially given the focus on finding matches in secondary data sources.

In this study, the research team devoted significant effort towards collecting alternate names for firms—or "aliases"—and entering these into the BLD reference database. All known aliases for a firm were run though an algorithm in SAS, which returned the top ten matches for each alias from the secondary database (both the NETS and the QCEW, separately). Based on the names and addresses in the BLD and the names and addresses in the matched NETS and QCEW records, the research team double-coded the algorithm's results as matches, potential matches or non-matches. The team then conducted double-coded additional research on the potential matches, using extensive web-based searches to either confirm or deny matches.

It is critical to note that the success of matching reference companies to secondary data sources is also influenced by the structure of the secondary data sources. The database structure of the NETS, for example, is far more conducive to matching and for identifying firms. The NETS has a single file of establishments for all years, so the matching algorithm only had to be run once. In

---

8 The algorithm relied on COMPGED function in SAS, which in turn is based on the Levenshtein distance between two string variables. The algorithm calculated the distance between the company name from the reference list (the BLD) and each company name in the secondary list (the QCEW or the NETS). The algorithm then sorted the numerical distances from smallest to largest, and returned the ten companies with the smallest numerical distances. Tests showed that this method of matching was faster than manual searching (i.e. searching for key words) and returned a higher number of matches.
contrast, the QCEW is comprised of yearly snapshots. Resources limited the research team to running the algorithm on only seven of the twenty years of QCEW data, though BLD reference data suggests this spread should be small enough to capture all firms in our study. Still, the remaining years will be examined in the future.

Validating Records from the QCEW and the NETS: While the BLD reference list contains start-up dates, locations, and entrepreneurial status, these don't necessarily match the information found in secondary data sources. The study's approach limited companies based on information found in the respective data sources. Thus after each reference firm is matched to a record in a secondary source, matches were removed if the secondary source included information indicating one of the following: first, that the first address in the database was not the company's first; second, that the company's first address was outside the core Triangle area, and; third, that the entrepreneurial company was in fact not entrepreneurial (i.e. it was part of a corporate hierarchy).

First, while the NETS contains a field for a firm's founding date, the QCEW does not. For consistency, the first year a company was listed in either the NETS or the QCEW was approximated to be its respective founding year. Since the QCEW files on hand for this research cover the 1990-2010 time period and NETS cover the 1989-2010 time period, companies that existed in 1990 in the QCEW and 1989 and 1990 in the NETS were dropped to ensure records reflected the company's first year of operation and covered the same time period.9

Second, street addresses from the QCEW were geocoded using the Texas A&M Geoservices online batch geocoding software.10 The NETS data contains latitude and longitudes. Locations outside the Triangle's core three counties were excluded.

Third, the NETS and the QCEW both have ways of identifying the corporate hierarchy of each establishment, though the NETS is far more robust. Companies that were found to have corporate hierarchies in 1991 were excluded.

Comparing the NETS and the QCEW: The study is designed to mimic what a researcher might be confronted with if she relied on a single data source for her research. Thus the QCEW and NETS are first assessed separately, and then compared against each other. Key indicators of the entrepreneurial environment in the Triangle's biomedical industry include: the number of companies from the reference database found in each secondary database, the initial employment levels found in the secondary databases, and the founding years in the secondary databases. The research first presents descriptive statistics for each of these descriptors for the QCEW and the NETS.

First, the number of companies found in each database is compared. A McNemar test assesses whether the number of companies included in the QCEW is significantly different from the number included in the NETS. Given the broader inclusion guidelines of the NETS, more reference firms are expected to be found in the NETS than in the QCEW.

---

9 It is possible that a company found in the QCEW had covered workers in the 1980s, had none in 1990, and then again had covered workers in subsequent years, leading to a false identification of an entrepreneurial firm. This is highly unlikely, especially given the few companies identified in the QCEW as start-ups in the early 1990s.

10 Available at http://geoservices.tamu.edu/Services/Geocode/
Next, initial employment rates are compared. A Wilcoxon signed rank sum test assesses whether there is a significant difference between the initial employment rates in the QCEW and the NETS in firms that are found in both databases. Given inclusion guidelines, the NETS is expected to have higher average employment rates than the QCEW.

The research includes a key adjustment to the NETS employment data. In their comparison of NETS and QCEW employment data, (Neumark et al., 2005) subtracted one employee per company. While it is not necessarily the case that these entrepreneurs do not receive wage labor, previous research suggests it is a likely (Kunkle, 2011; Neumark et al., 2005). While they were forced to rely on aggregate data and proxy entrepreneurial employment, the extensive background data in the BLD reference list allows for a precise subtraction of entrepreneurs for each company. These adjusted NETS employment numbers are then compared against the QCEW employment numbers, and Wilcoxon signed rank sum test assesses whether there is a significant difference between this adjusted NETS employment rates in the QCEW. Removing entrepreneurial employment from the NETS records should remove the majority of the difference between the NETS and the QCEW, as it removes a key purposeful difference between the databases.

A similar entrepreneurial adjustment (i.e. subtracting known entrepreneurs for all NETS firms) is conducted on all firms, as part of an adjustment to the number of firms found in each database. Firms with zero or negative entrepreneurial-adjusted employment are removed the NETS matches, and the adjusted number of NETS matches is compared against QCEW matches. This is expected to remove firms that would not be in the QCEW, and remove the majority of the difference between the NETS and the QCEW, as it removes a key purposeful difference between the databases.

Differences in the founding years are assessed in two ways. First, a Wilcoxon signed rank sum test assesses whether there is a significant difference in the founding years. Next, trend lines for both databases are compared. On average, NETS companies are expected to have been founded in earlier years than QCEW companies, though trends in the region should be similar.

Finally, kernel density maps assess spatial concentrations of the QCEW and NETS entrepreneurial environments. The NETS is expected to show greater diffusion than the QCEW, as pre-incorporation, pre-covered worker addresses are more likely to be in residential locations than addresses for incorporated firms with covered workers.

Results

Number of Entrepreneurial Firms: The BLD reference list of biomedical entrepreneurial start-ups contains 647 firms. Of those, 478 (73.8 percent) companies were matched to records in the NETS, the QCEW, or both. As expected, a larger number of entrepreneurial companies were matched to records in the NETS than in the QCEW. Of the 647 biomedical firms, 417 (64.5 percent) were matched to the NETS, whereas only 261 (or 40.3 percent) were matched to the QCEW (see Table 1). Of the 261 matches to the QCEW, just 61 (23.7 percent) were unique to the QCEW, whereas 217 of 417 (52.0 percent) were unique to the NETS. Just 200 of the 647 firms (30.9 percent) were found in both databases. The McNemar chi2 statistic suggests that the proportions of the BLD reference database found in the NETS and the QCEW are statistically significant from one another.
Table 1: Entrepreneurial Start-up Firms in the QCEW and the NETS

<table>
<thead>
<tr>
<th></th>
<th>All NETS</th>
<th>All QCEW</th>
<th>Adjusted NETS</th>
<th>Adjusted QCEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firms</td>
<td>417</td>
<td>261</td>
<td>292</td>
<td>261</td>
</tr>
<tr>
<td>McNemar's chi²(1)</td>
<td>87.54</td>
<td>4.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exact McNemar</td>
<td>0.0000</td>
<td>0.0501</td>
<td></td>
<td></td>
</tr>
<tr>
<td>significance probability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Of the 417 firms matched to NETS records, 125 have entrepreneurial employment that was either equivalent to or equal to initial employment levels. Removing these firms brings the adjusted NETS matches down to 292. The McNemar chi² comparison is statistically insignificant (though only just), suggesting that the difference in the number of firms matched using adjusted numbers is not statistically significantly different.

Employment: On average, the firms in the NETS had higher levels of employment than the firms in the QCEW. The 417 firms in the NETS had, on average, 0.9 more employees per firm than the 263 firms in the QCEW (6.2 vs. 5.3, respectively; see Table 2).

Table 2: Employment Statistics, QCEW and NETS

<table>
<thead>
<tr>
<th></th>
<th>NETS</th>
<th>QCEW</th>
<th>Matched NETS</th>
<th>Matched QCEW</th>
<th>Adjusted NETS</th>
<th>Adjusted QCEW</th>
</tr>
</thead>
<tbody>
<tr>
<td>Firms</td>
<td>417</td>
<td>261</td>
<td>200</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Average employment</td>
<td>6.2</td>
<td>5.3</td>
<td>7.8</td>
<td>4.8</td>
<td>6.1</td>
<td>4.8</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>9.5</td>
<td>8.7</td>
<td>11.6</td>
<td>8.1</td>
<td>11.6</td>
<td>8.1</td>
</tr>
<tr>
<td>Minimum</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Maximum</td>
<td>104</td>
<td>89</td>
<td>104</td>
<td>89</td>
<td>104</td>
<td>89</td>
</tr>
<tr>
<td>Wilcoxon signed- rank test z-statistic</td>
<td>N/A</td>
<td>-6.421</td>
<td>-0.512</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The 200 firms in both the NETS and the QCEW have average employment of 7.8 and 4.8, respectively. Before adjusting for entrepreneurship, the 200 NETS and QCEW firms contained average initial employment levels of 7.8 and 4.8 per firm, respectively. A Wilcoxon signed-rank test suggests there is a statistically significant difference between these employment rates. After adjusting for entrepreneurial employment, the NETS average dropped from 7.8 to 6.1, removing over 57 percent of the initial difference between the QCEW and NETS matched average employment rates. A Wilcoxon signed-rank test suggests there is not a statistically significant difference between employment in the NETS and employment in the QCEW.
Firm Births Over Time: Of the 200 entrepreneurial companies matched to records in both the QCEW and the NETS, only 22 were recorded as being founded in the same year. Of the remaining 178 establishments, 159 were founded earlier in the NETS and 19 were founded earlier in the QCEW. The mean difference in founding years was 1.5 years earlier in the NETS. A Wilcoxon signed-rank test suggests there is a statistically significant difference between the founding years in the NETS and in the QCEW (z-statistic of 8.690, p=0.000).

Figure 3: Number of Firm Births per Year, QCEW and NETS, and Trend Lines

The distribution of firm births over time resulted in different entrepreneurial birth trends in the region. Figure 3 shows the number of firms founded per year in the QCEW and the NETS, respectively, and their trend lines. While the two databases had similar starting points, over time the number of firms in the NETS has increased at a much faster rate than those in the QCEW.

Location: Figures 4 and 5 show the locations of all 261 QCEW firms and 417 NETS firms in the three county Triangle region, respectively, and their locational relationships to the Research Triangle Park, Duke University, UNC, and North Carolina State University. From these maps it appears that firms in both the QCEW and the NETS center around and spread out from the Park, but that the NETS has many more firms.

allows for hot spot smoothing, show that both the QCEW and the NETS have peak entrepreneurial firm density centered around the Research Triangle Park, though the concentration is stronger in the NETS than in the QCEW.

Yet, kernel density maps built a bandwidth that allowed for more distinct hot spots show that while the NETS has a stronger concentration of firms in the core of the region, NETS firms are also more diffused around the region, with concentrations of firms farther out from the peak Research Triangle Park area (see Figures 8 and 9).
Figure 4: Biomedical Industry Firms in the QCEW, with Research Triangle Park and Three Universities, 1991-2010

Figure 5: Biomedical Industry Firms in the NETS, with Research Triangle Park and Three Universities, 1991-2010
Figure 6: Kernel Density Map of Biomedical Industry Firms in the QCEW, High Bandwidth (0.5)

Figure 7: Kernel Density Map of Biomedical Industry Firms in the NETS, High Bandwidth (0.5)
Figure 8: Kernel Density Map of Biomedical Industry Firms in the QCEW, Low Bandwidth (0.05)

Figure 9: Kernel Density Map of Biomedical Industry Firms in the NETS, Low Bandwidth (0.05)
Kernel density maps allow for a visual exploration of differences in the concentration or "hot spots" of firms in the Triangle. Figures 6 and 7, both developed using a high bandwidth that

Finally, of the 200 companies in both the NETS and the QCEW, 60 (30 percent) had records within 100 meters of each other, and 118 (59 percent) had records within 1600 meters of each other. On average, matched companies were 3976 meters apart, with a standard deviation of 6561 meters and a range of 2.7 to 37172 meters. Seventy-one percent of companies (142 of 200) were within the same city, and 84.5 percent (169 of 200) were within the same county.

**Discussion**

The results suggest there is substantial variation between records contained in the NETS and the QCEW. Relying on records from just one data source give noticeably different results for the Triangle region, with the NETS giving the impression of a larger entrepreneurial environment diffused across a larger portion of the three-county region, with higher average employment levels. Firms appear older in the NETS than in the QCEW, and the rate of firm births is increasing faster in the NETS than in the QCEW. All of these findings, with the exception of the larger growth rate of firm births in the NETS, are consistent with the literature.

Yet, the study also suggests that purposeful differences may drive much of these differences. For example, while there is a statistically significant difference between the number of reference entrepreneurial firms found in the NETS and the QCEW (417 vs. 261), this difference drops dramatically when NETS firms consisting solely of entrepreneurs at the time of their founding are excluded (292 vs. 261), and the statistical significance of the difference disappears. Similarly, while the NETS and the QCEW initial employment rates in the 200 matched firms are statistically significantly different, almost 60 percent of the difference and the significance of the difference vanishes when entrepreneurial employment is removed. Spatially, while the NETS is more diffused across the region, an examination of the addresses in the NETS suggests that the greater diffusion may be the result of firms using residential addresses, which may represent unincorporated businesses run out of an entrepreneur's home.

Still, the research results also point to several additional areas of future research:

First, the difference in founding years complicates comparisons between the databases. For example, it is unclear if the spatial diffusion of NETS data is due to the inclusion of firms run out of home offices that were and never will be eligible for the QCEW, or firms that were ineligible at their founding but later moved to a more central area. To control for this, future versions of this research will also include an analysis of employment and spatial differences for in the same years.

Second, it is possible that comparisons between the NETS and the QCEW mask differences in match rates that would become apparent with more detailed explanation of industry. The holistic conceptualization of the biomedical industry may have masked, for example, greater match rates for pharmaceutical entrepreneurial companies than for drug development consulting companies. Differences in covered workers across detailed industry may occur, and prior research on food establishments has shown that match rates may vary across source (see, for example, Fleischhacker et al., 2012; Liese et al., 2010, 2013; Ma et al., 2013; Powell et al., 2011).
Third, the research will assess the accuracy of the latitudes and longitudes provided in the NETS. The NETS occasionally defaults to block group, census tract, or zip code centroids. While this study relied on given latitudes and longitudes to best represent the approach the most researchers take (see Liese et al., 2013), an assessment of the spatial accuracy between the QCEW and the NETS should also include an analysis using freshly geocoded addresses.

Fourth, further analysis into the differential firm birth trends is needed. It is unclear if the NTES is simply getting better at capturing firms, or if the BLD reference database is capturing more informal firms over time.

Even without these additional areas of research, the results reinforce the idea that the QCEW may not be appropriate for studies of entrepreneurship (Buss, 1995; Kunkle, 2011), if the goal of the entrepreneurial research is to gain a better understanding of an entrepreneurial environment that includes unincorporated firms and/or firms comprised solely of unpaid workers. The role that entrepreneurial employment plays in both employment differences and firm match differences suggests that the NETS does a much better job at capturing entrepreneurial firms and their employment rates, and that purposeful differences—not problems with accuracy—drive differences between the NETS and the QCEW. More generally, the research suggests that the NETS and the QCEW should not be considered perfect substitutes, and that researchers must think through purposeful differences between any alternative data source and their "gold-standard equivalent" when designing their research.

Finally, the research suggests that future studies of alternative database accuracy should attempt to differentiate between purposeful and accidental differences, either as thought exercises or by making adjustments to their comparisons, as this paper has done. While this research focused on two very specific databases, the approach taken here—exploring in depth the differences in inclusion guidelines, hypothesizing what key differences may be, and testing these differences in ways that exploit externally-gathered information on entrepreneurial firms—can be carried to other assessments.
References


Liese, A. D., Barnes, T. L., Lamichhane, A. P., Hibbert, J. D., Colabianchi, N., & Lawson, A. B. (2013). Characterizing the food retail environment: Impact of count, type, and geospatial...


The Southern California Housing Bubble: Neighborhood Level Measures and Economic Implications

by Johannes Moenius | APRIL 1ST, 2014

Introduction

Southern California has seen a substantial increase and subsequent drop in home values. The increase during the upswing afforded home owners the possibility of increased consumption through borrowing against their perceived new wealth. During the downturn, this development was reversed, often leaving those who had borrowed against their houses during the upswing with no or negative equity in their homes. Neighborhoods with high housing turnover became less desirable causing further subsequent price drops. Property tax revenues first rose and then fell, and the banks have had to absorb an increasing mountain of mortgages in default.

As these events did not affect all locations in Southern California equally, we analyze their spatial distribution. We propose expanding our preliminary 2009 study (http://isea.redlands.edu/analysis/2009/04/17/socal-housing-market/) and use individual transaction data on single family homes to construct and improve on seven indicators per square-mile and quarter:

1. Value at risk
2. Standardized home price
3. Housing unit turnover
4. Equity at time of purchase
5. Consumption possibilities out of housing wealth
6. Property tax income as determined at the time of purchase
7. Mortgage default risk
Change in Value at Risk / Price Development in Southern California

We measure the change in value at risk as the changes in average as well as total amount of spending on single family homes by square mile, regardless of home size. We use hedonic regressions to construct (counterfactual) standardized home prices per square mile.

Southern California has experienced a dramatic development of average value at risk per home sold (see animation: http://www.youtube.com/watch?v=k8hBHbLtL_A). Real average value at risk more than doubled on average, but certain neighborhoods saw average value at risk more than triple. This development was by no means uniform as some neighborhoods substantially rose more than others. Home prices in neighborhoods that were more desirable at the outset both rose and fell less than those in less desirable neighborhoods. However, the rise and fall were asymmetric, indicating a substantial redistribution of housing wealth across locations.

1998-2000 SoCal average price per square foot
2006 SoCal average price per square foot

2008-2009 / Q4-Q1 SoCal average price per square foot

Housing Unit Turnover

We use housing turnover as a proxy for neighborhood stability and measure it with the number of transactions from 1998 to 2014 relative to the number of housing units in the year 2000 per square mile. Data for the Inland Empire indicates that in some neighborhoods the number of transactions during the analyzed decade has been more than 50% of the number of units, while others saw only minimal turnover.

Equity at Time of Purchase

The down-payment new homeowners offer at the time of purchase determines how much buffer the banks have in case of mortgage default. Once home prices fall banks may suffer
losses if down-payments are lower than the market value loss on the property in default. During the bubble, down-payments have been surprisingly low in Southern California, with negative average equity for many neighborhoods at the time of purchase for homes reported to have been financed with at least one mortgage.

In the animation at http://www.youtube.com/watch?v=Cf3oR3ObslI, red squares indicate negative equity shares (= equity as a percentage of sale price) at the time of purchase, green squares positive equity shares. After the bubble started bursting, banks required higher down-payments. Two features stand out: negative equity shares are concentrated in those areas that exhibited the strongest growth and later decline in house prices (namely the Inland Empire), and average equity shares during the early downturn were lower in most areas than the decline in home prices.

**Consumption Possibilities out of Housing Wealth**

Increasing housing wealth implies larger consumption possibilities. Wealth effects were hugely positive during the bubble upswing. Anyone who purchased a house even with negative equity saw the red ink eroded and replaced with continued perceived wealth increases. All of Southern California benefitted tremendously from this effect.

The animation at http://www.youtube.com/watch?v=bGC5FRYwlSE shows the dramatic increase in equity in each neighborhood. It also shows how equity was slashed even faster during the downturn and the amount of negative equity increased. This process has happened the faster and the more pronounced, the more prices fell.

**Property Tax Income as Determined at the Time of Purchase**

Property taxes in California are generally determined at the time of purchase of a property, can only rise by a limited amount (Proposition 13), but can be adjusted downward if the property value shrinks (Proposition 8). The combination of the two propositions poses substantial fiscal risk to local municipalities and ultimately to the State of California, who has to cover any shortfalls of property tax income at least for schools. In some areas, average property taxes of newly sold homes halved on average from the peak to the first quarter of 2009. This implies also lower property tax payments for homes that have been sold in the late bubble years due to Proposition 8.
2006 SoCal property tax revenue

2009 SoCal property tax revenue

Default Risk

Mortgage default risk increases if the amount of negative equity in a home increases. The more houses in a neighborhood are deep in the red, the more likely will this neighborhood see additional mortgage defaults, foreclosures and thus additional decreases in neighborhood stability and home values. We present a simple risk measure, the skewness of negative equity. In the map below, dark green areas are the relatively safest, while red areas are the most likely to see defaults in the future.
Summary

We suggest constructing indicators on the neighborhood level from individual transaction data to gauge developments and risks in real estate markets. The analysis indicates that economic distress on several dimensions is highest in areas where house-prices have risen and fallen the most. Even as the national economy and housing market recovers, these areas could suffer prolonged economic weakness.

Data-sources: ACS, Census, ESRI, Dataquick
Examining Intraurban Migration in the Twin Cities Metropolitan Area using Parcel Data

Shipeng Sun

Institute on the Environment
University of Minnesota, Twin Cities
1954 Buford Avenue
Saint Paul, MN 55108
Phone: (612) 232-3455
Email: sunx0170@umn.edu, sunsp.gis@gmail.com
Abstract

Intraurban migration, i.e., residential relocations within a metropolitan area, is an important aspect of urban and housing studies, particularly for regional housing market, transportation planning, urban residential structure, and behavior-based micro urban modeling. Intraurban migration is tightly embedded in regional housing market and simultaneously creates and absorbs housing vacancies. Households' relocation decisions are manifestations of housing preferences, employment opportunities, traffic conditions, and residential structure. The ethnicity and social status of those migrants also gradually yet forcefully leads to the rearrangement of urban space. Moreover, intraurban migration studies offer a valuable empirical benchmark to micro-urban models like those agent-based for their specification, calibration and validation.

The paucity of data on the migration choices of individuals, however, remains a critical challenge in understanding intraurban migration. While migration evinces clear patterns such as suburbanization, gentrification, or decline when examined at gross temporal and spatial scales, our understanding of migration at the scales of individuals is limited by the dearth of public data available on movements of individual households. A common approach to measuring intraurban migration is surveying individuals regarding their recent moves and then reporting on them over large enumeration units. These sources range from travel surveys to general instruments such as the American Community Survey (ACS), the American Housing Survey (AHS), the Current Population Survey (CPS), and the Public-Use Microdata Samples (PUMS). They offer good
information about intraurban migration in general but lack the spatial resolution necessary to analyze individual moves at subregional scales due to spatial aggregation. Another common approach to measuring intraurban migration is to gather data on specific households or houses in an area. Directly surveying migrants is a good way to understand their home seeking behavior, but this approach is expensive and typically reaches only a small subset of migrants. A related approach is using home sales data to capture attributes of specific homeowners, but these data usually say little about the search and migration behavior of specific individuals and often include false moves resulted from speculation. Overall, data on specific households and houses offer spatial specificity not found in aggregate data noted above, but their use is not without challenges.

Using dig data techniques, we developed a novel form of information on household intraurban migration to address key data challenges, namely migration chains from land parcel data for an entire region. A migration chain establishes linked pairs of moves, each defined by a household that leaves a property and one that moves into the just-vacated property. Parcel data are suited to this task when they encompass all home ownership for a specific area; in the Twin Cities, for example, these data describe over one million lots each year. This research utilizes the annual regional parcel dataset in the Twin Cities compiled and managed by the regional government, the Metropolitan Council, spanning the seven counties of Anoka, Carver, Dakota, Hennepin, Ramsey, Scott, and Washington. Relevant information includes owner’s name and date last sold; other data vary by jurisdiction, such as square footage of houses and their lots or dwelling type. More specifically, we developed migration chains for the Twin Cities by comparing
the owners of a parcel across years, detecting valid owner changes and matching owners across years. We weeded out transactions, such as speculation and bank sales, that represent ownership change without a household move. We also left out condominiums and apartments given that many are not owner occupied (so renters are not included). We developed software that embodied a multipart strategy to deal with variations and errors in names. For example, names with various abbreviations and forms are automatically recognized and combined. Household name changes via marriage and heritage are also handled to avoid the inclusion of false moves. Overall, this method produces a relatively exhaustive and fine-scaled delineation of the residential moves within the metropolitan area across multiple years.

With such intraurban migration information extracted from parcel dataset with big data techniques, we examined the Twin Cities regional housing market from the unique perspective of migration. First, we calculated the housing turnover and migration rate from the number of homeownership changes. Different types of new and demolished houses are also categorized and counted to illustrate their varying up-and-down trends before and after the 2008 financial crisis. At the individual level, the extraction of housing vacancy chains and their length distribution and comparison to other studies shed light on how these features of housing relate to new housing construction and homeownership rates. Move distances and directions mapped at multiple scales illustrate the complexity of the spatial dynamics of intraurban migration while confirming the persistence of traditional patterns and mathematical models of migration. Finally, exploring the changes of structural and neighborhood characteristics after migration reveals features of social mobility, or how migration and socioeconomic status interact.
In sum, regional parcel data in combination with big data techniques can be used to advance our knowledge of housing and intraurban migration patterns, and to lay solid behavioral foundation for the further exploration of complex housing locational decision process that is critical to understand many aspects of regional dynamics.
Modeling Urban Capacity with Public Data: Helping Realize Universal Prekindergarten in New York City

Richard Dunks, Ravi Shroff, Jeongki Lim, Miguel Castro, Haozhe Wang
{rad416, rs4606, jl2684, mac1077, hw1067}@nyu.edu
New York University
Center for Urban Science & Progress
July 15, 2014

Figure 1: Distribution of four year olds in New York City based on data from US Bureau of the Census (left) and the distribution of publicly funded prekindergarten programs (right)

Abstract

Using public data from the US Bureau of the Census and the New York City Department of Education, we consider the problem of identifying locations in New York City where four year olds are potentially underserved by available capacity in city-funded prekindergarten programs. We were able to create a map of four year old children in New York City and the available prekindergarten capacity in both public schools and publicly-funded community-based organizations. We implement a random allocation algorithm to identify and map underserved locations. As part of initial work in this area, our model incorporates a rough travel distance measurement and has potential for being improved with better travel estimates and a more nuanced modeling of travel flows in New York City, as well as more sophisticated techniques for estimating the location of four year olds in New York City using administrative data on residential tax lots.
1 Introduction

The implementation of a publicly funded universal full-day prekindergarten (pre-K) program has become an important topic in New York City politics after having figured prominently in the 2013 campaign for Mayor of New York City. Studies have shown numerous benefits from high quality pre-K instruction, including increased cognitive abilities, higher test scores in the short term, and access to higher paying jobs in the long term (see Gormley, 2004 for a fuller discussion of these effects).

Currently, public pre-K programs in New York City are offered in public schools and community based organizations (CBOs) receiving public funds. Programs are half day (two hours and 30 minutes of instruction) or full day (six hours and 20 minutes of instruction). While admission to CBO programs is first come, first served, admission to public school programs requires parents to apply by the end of the previous school year and select up to 12 schools for their child to attend. Admission decisions are based on a priority list giving first priority to students who live in the zone for that particular school, then students living in the school district but without a pre-K site in their zone, students living in the same school district but not in the same school zone as the school, students living in the same borough as the school, and finally out-of-borough students, in that order. At each point, having a sibling already attending the school increases the prioritization of an applicant for a seat at that particular school (Department of Education, 2014).

The NYCDOE estimates there are 73,250 children who require access to full-day prekindergarten programs. This number is derived by taking the 81,748 children enrolled in kindergarten and subtracting the estimated 8,498 children who will likely enroll in private pre-K. There are currently 58,528 pre-K seats available in public schools and in publicly-sponsored pre-K programs run by CBOs under contract to either the NYC Department of Education (NYCDOE) or the NYC Administration for Children’s Services (NYCACS), which manages pre-K for low-income children with support from the federal government. This includes 26,364 half-day seats and 32,164 full-day seats. Of the currently available seats, 23,671 are in public schools and 34,857 are in CBOs, although the full-day seats are almost evenly distributed between public schools and CBOs (Office of the Mayor, 2014).

To meet the gap between children requiring full time pre-K and the current capacity (approximately 41,000 children), the NYCDOE plans to expand the pre-K system by 23,640 in the 2014-2015 school year, split evenly between the conversion of part time seats to full time seats and an increase in the number of available seats, and 17,446 in the 2015-2016 school year, with most of the increase coming in the conversion of half-day seats to full time seats (Office of the Mayor, 2014).

The purpose of this paper is to use publicly available data to model the distribution of
four year olds in New York City and identify areas currently underserved by existing pre-K capacity. Our intention is to develop a more sophisticated approach to modeling capacity that can better inform the decision-making process around increasing access to high-quality pre-K, particularly in areas where the added benefit of pre-K instruction can matter most. Our work is preliminary and we note areas where both the available data and methodology could be improved.

2 Data Sources

We use publicly available data from the US Bureau of the Census, the New York City Department of Education, and PediaCities, a platform to “curate, organize, and link data about cities”\(^1\) in order to identify areas currently underserved by existing pre-K capacity. The combination of two open data sets from different governmental organizations (federal and city) present unique challenges to the task of analyzing public data for the public good. We use standard geospatial processing techniques to combine and visualize the data, as well as standard statistical methods for estimating the distribution of four-year-old children throughout the city.

2.1 Census Data

The number and location of four year olds was derived from the 2012 5-year aggregated American Community Survey (ACS). The ACS groups the population into various age bands for reporting purposes. To estimate the number of four year olds living in a particular census tract, the population under the age of 5 (the age band in which the Census reports four year olds) was assumed to be evenly distributed among the 5 years encompassed by children ages 0 - 4. The population number was divided by 5 to arrive at the estimate of four year-olds for a given census tract. This yields an estimate of 105,410 four year olds in New York City\(^2\).

2.2 Prekindergarten Location and Capacity

Data on schools was gathered from both the NYCDOE open data portal\(^3\) and PediaCities. Each pre-K site is identified by a six digit site ID, with the first two digits indicating the school district number, the third digit a letter indicating the borough, and the remaining three digits a unique identifier for the site. From the NYCDOE data we were able to collect capacity information for the public school (PS) sites while PediaCities provided locations for both the PS and CBO sites. For the 2013-2014 school year, NYCDOE lists 1,406 pre-K sites. Of these, 29 public schools listed as having provided pre-K at some point in the past no longer have

---

\(^1\)Information about PediaCities is available from their website, [http://www.pediacities.com/](http://www.pediacities.com/)

\(^2\)The disparity between this number and the number used in Office of the Mayor, 2014 is discussed below

\(^3\)Available at [http://nycdoe.pediacities.com/](http://nycdoe.pediacities.com/)
pre-K seats. One school, PS 051 Elias Howe, hasn’t had pre-K seats since the 2010-2011 school year, but will have seats for the 2014-2015 school year. To keep the analysis consistent, this school was excluded from this analysis. Among the community-based organizations (CBOs), 6 sites were listed without site IDs and appear to no longer host pre-K seats. These have also been excluded from the analysis.

There are 34,857 CBO seats reported at approximately 850 sites (Office of the Mayor 2014). This matched the 855 CBOs listed in the PediaCities data. Of these 855 CBO sites, NYCDOE had readily available capacity data for only two sites, Baychester Academy and the Staten Island School of Civic Leadership with reported capacity of 36 and 18, respectively. Subtracting these 54 seats from the CBO total leaves 34,803 seats spread across 853 facilities, or an estimated 40.8 seats per facility. For the purposes of this analysis, each CBO for which the actual capacity is unknown has been assigned a capacity of 41.

For the purposes of this analysis, equal-sized hexagonal polygons 650 meters across were used to divide the city. This created a common spatial unit of analysis among the various geographical divisions, allowing for data to be broken down into smaller units and then aggregated as necessary to compare capacity within various political, administrative, and other delineations.

3 Methods and Tools
We divided this task into two main phases. In the first phase, we mapped the locations of both four year olds in New York City and the pre-K locations. We also added attributes for the pre-K sites based on available information. In the second phase, we applied an allocation algorithm to model the availability of pre-K seats for the population of four year olds we mapped in the first phase.

3.1 Mapping Distribution of Four Year Olds
The US Bureau of the Census releases demographic data aggregated into various statistical subdivisions that follow political and administrative boundaries. The common unit of analysis is the census tract, an area with approximately 1,200 to 8,000 people that can vary in size depending on the population density of a given area. In addition to tracts, the Census releases data in blocks, block groups, zip codes, Census designated places (CDPs), counties, metropolitan and micropolitan statistical areas, states, tribal areas, and the country as a whole. For the purposes of this analysis, data for census tracts were used to estimate the population of four year olds in New York City.

In order to create a common spatial unit for

---

4The NYCDOE did not release comprehensive CBO capacity information for the 2013-2014 school year in a machine readable format. An audit of CBO capacity would require going through PDF directory listing intended for parents to select a pre-K to match sites to ensure they appear in the current listing of available programs.
analyzing the population of four year olds and the closely located pre-K capacity over a consistent area, we used a mesh of 2,930 equal-sized hexagons measuring approximately 650 meters across. The size was chosen to provide a unit of analysis that was small enough to provide an estimate of a smaller area than that generally covered by a census tract without being too small that it created a computational task that couldn’t have been accomplished on our available hardware. We then employed a technique known as dasymetric mapping to estimate the number of four year olds living in a particular hexagon.

Dasymetric mapping is a technique for disaggregating spatial data into smaller units of analysis based on ancillary information. The technique was first described in 1911 by Benjamin Semenov-Tian-Shansky and employed in his 1923 “Dasymetric Map of European Russia” that used the land use categories to estimate population densities (Petrov, 2012). Dasymetric mapping has since been developed to employ various techniques to match data between spatially mismatched delineations (Mennis, 2009; Zandbergen and Ignizio, 2010).

In this analysis, we employed a basic areal weighting that takes into account the area overlap between the two spatial units, in this case the census tract and the hexagon. Assuming the demographics of a census tract are evenly distributed across the tract, a portion of the demographic characteristics equal to the portion of the tract area covered by the hexagon is then added to the hexagon. For example, if a hexagon completely overlapped the census tract, the hexagon included the entire estimated population of four year olds. If the hexagon only overlapped 50% of the census tract, the hexagon received only half the four year old population. The number of four year olds from each of the constituent census tracts was then summed and rounded to the nearest integer to provide a number of four year olds in the hexagon.

The statistical delineations used by the Bureau of the Census follow administrative boundaries. In the case of New York City, this means that many of the 2,167 census tracts include a significant amount of water area as the boundaries between New York and New Jersey, as well as borders between the 5 boroughs, fall in the middle of local waterways. Assuming that most, if not all people live on land, we excluded the water area from the areal interpolation method described above and used census tracts clipped to the shoreline of New York City.

Using the locations of pre-K sites from the NYCDOE available from PediaCities, the number of public school (PS) and CBO pre-K sites were joined to each hexagon. The number of seats were summed for all sites located within a hexagon, separated by whether they were PS seats or CBO seats. No distinction was made between part time and full time seats. Thus a site with 18 morning seats

---

5We used 2010 census tract files prepared and released by the NYC Department of City Planning that exclude the water area from and available on their website at [http://www.nyc.gov/html/dcp/html/bytes/districts_download_metadata.shtml](http://www.nyc.gov/html/dcp/html/bytes/districts_download_metadata.shtml)
and 18 afternoon seats would have a capacity of 36. Each hexagon was also assigned to a particular school district. For those hexagons lying along district boundaries, the district that occupied 50% or greater of the hexagon area was assigned as the district for that hexagon.

To simulate the ability of parents to take their kids to nearby pre-Ks outside their hexagon, we created a list of nearby hexagons based on a simple distance measurement. A nearby hexagon was defined as a hexagon within 2,000 meters of the given hexagon. This simulates approximately 15 minutes of travel time. We chose this estimate in order to validate the approach with the intention of employing a more sophisticated approach, which we outline below.

### 3.2 Calculating Need

Our allocation algorithm (Algorithm 1) uses Monte Carlo methods to determine areas of New York City underserved in terms of pre-K program access. The input consists of the geographic distribution of four-year-olds and current PS and CBO pre-K capacities for each of the 2,930 hexagons described above. We simulated the ability of parents to take their children to nearby pre-K seats by creating for each hexagon $H$ a list of “nearby” hexagons, defined to be those hexagons overlapping a disc of fixed radius (the “travel distance”) centered at the given hexagon. We assume that although a child in $H$ may attend any CBO pre-K program in any nearby hexagon, he or she may only attend those public school pre-K programs in nearby hexagons in the same school district as $H$. This assumption is consistent with public school pre-K admissions criteria, which give strong preference to an applicant whose residence is in the same school district as the target school. Without detailed family information, we were unable to model the sibling-preference in PS prioritization.

Each hexagon $H$ has resident population of children $P_H$, current number of assigned public school and CBO students $S^{ps}_H$ and $S^{cbo}_H$ respectively, public school and CBO capacities $C^{ps}_H$ and $C^{cbo}_H$ respectively, a list of nearby hexagons $F_H = \{F_1, \ldots, F_{j_H}\}$ in the same

<table>
<thead>
<tr>
<th>Algorithm 1</th>
</tr>
</thead>
</table>
| **Input:** Hexagons $\mathcal{H} = \{H_n\}^{2930}_{n=1}$, $H_n = \{P_{H_n}, S^{ps}_{H_n}, S^{cbo}_{H_n}, C^{ps}_{H_n}, C^{cbo}_{H_n}, F_H, G_H\}$ 

Initialize parameters:
- $P_{H_n} = P_{0H_n} \{\forall n\}$
- $S^{ps}_{H_n} = 0 \{\forall n\}$
- $S^{cbo}_{H_n} = 0 \{\forall n\}$
- $\mathcal{X} = \{H \in \mathcal{H} \text{ s.t. } H \text{ usable}\}$

while \(\text{length}(\mathcal{X}) \neq 0\) do
  randomly choose $H \in \mathcal{X}$
  randomly choose non-full neighbor $K$ of $H$ from either $F_H$ or $G_H$
  if $K$ chosen from $F_H$ then
    $S^{ps}_K = S^{ps}_K + 1$
  else
    $S^{cbo}_K = S^{cbo}_K + 1$
  end if
  $P_H = P_H - 1$
  $\mathcal{X} = \{H \in \mathcal{H} \text{ s.t. } H \text{ usable}\}$
end while

**Output:** Pairs $\{(n, o_n)\}^{2930}_{n=1}$ where $o_n = \frac{P_{0n}}{P_{H_n}}$ if $P_{H_n} \neq 0$, and $o_n = -1$ otherwise
school district as $H$, and a list of all nearby hexagons $G_H = \{G_1, \ldots, G_{kh}\}$ (so $F_H \subseteq G_H$). Initially, $S_{ps}^H$ and $S_{cbo}^H$ are set to zero and $P_H$ is set to $P_0^H$, the initial resident population.

We say any hexagon $K$ is a non-full neighbor of $H$ if $K \in F_H$ and $S_{ps}^K < C_{ps}^K$ or $K \in G_H$ and $S_{cbo}^K < C_{cbo}^K$. We say $H$ is usable if it has at least one child ($P_H > 0$), and at least one non-full neighbor.

The algorithm then works as follows. While there exists at least one usable hexagon, randomly choose such a usable $H$ and randomly choose a non-full nearby hexagon $K$ from either $F_H$ or $G_H$. Decrement $H$'s resident population $P_H$ by 1 and if $K$ was chosen from $F_H$, increment $K$'s assigned public school students $S_{ps}^K$ by 1, otherwise increment $K$’s assigned CBO students $S_{cbo}^K$ by 1.

When the algorithm terminates, for each hexagon $H$ compute the output statistic $P_H / P_0^H$, the percentage of resident children in $H$ that were unable to be allocated to a pre-K spot. We chose to run the algorithm six times due to computational limitations and average the output statistics. Finally, we visualized the averaged output statistic by creating a map where each hexagon was shaded from green to red based on the output statistic (red means all children were unallocated and green means all children were allocated to pre-K seats). Hexagons that did not possess children to begin with are shaded white. For work related to this approach, see Holmberg, et al., 1999.

3.3 Software Tools
We parsed the available text files using Python scripts before uploading them to a PostgresQL database hosted on the Amazon Web Services cloud computing platform. We used the PostGIS spatial extension to perform the areal weighting of census tracts to hexagons and the dasymetric spatial join of demographic data to hexagons. We coded the simulation algorithm in Python to run on data exported from the PostgreSQL database, joining results back to the database at the conclusion of the program. We visualized the data using QGIS, an open-source geospatial information system (GIS). All tools used in this analysis are open-source and freely available online.

4 Results and Discussion
Our analysis shown in Figure 2 identified areas in all 5 boroughs, particularly the neighborhoods in Sunset Park in Brooklyn (red box lower left), Corona (middle right) and Far Rockaway (bottom right) in Queens, the Upper East Side and the Upper West Side in Manhattan (middle left), and the North Bronx (top right). In these neighborhoods, we estimate there is a high percentage of four year olds who aren’t able to find a pre-K slot nearby based on 2013-2014 capacity.

Figure 3 shows a similar analysis looking at the estimated number of four year olds unable to find a spot (as opposed to the percentages shown in Figure 2). While the areas
of Sunset Park in Brooklyn and Corona in Queens are underserved by percentage, Sunset Park shows a higher number of four year olds who likely won’t be able to find a pre-K spot nearby. The contrast between Sunset Park and Manhattan areas is interesting as Sunset Park is known as a low-income area with a number of immigrants while the underserved areas of Manhattan are known as high-income areas where parents likely have access to private schools rather than rely on publicly funded options. This suggests that impediments to truly universal prekindergarten exist at both ends of the socio-economic spectrum.

5 Limitations and Future Work

Ultimately, the reliability of this approach rests on the accuracy and completeness of the data upon which it is built. Not having accurate CBO capacity information prevents us from making an accurate estimation of the available capacity throughout New York City. Having this data easily available in a machine readable format would not only enable work such as ours, but also encourage entrepreneurs to develop web and mobile applications to provide parents with this critical information.

Our population estimate of 105,410 estimates all four year olds in New York City whether they are enrolled in public school or
not. The figure of 81,748 four year olds used in Office of the Mayor (2014) is based on the number of five year olds enrolled in public kindergarten. The difference of 23,662 is possibly explained by the number of five year olds enrolled in private school; however the 2012 5-year ACS estimates this number to be 10,855. It’s unclear where this discrepancy comes from and suggests a considerable difference in how the NYCDOE and US Bureau of the Census estimate population in New York City.

Our use of a simple dasymetric modeling approach with areal weighting likely doesn’t provide the most accurate estimate of four year olds for a given area. This could be refined by using administrative data on the location of tax lots with residential units similar to work in Maantay, et al. 2007 and Maantay, et al. 2008, which used data from the NYC Primary Land Use Tax Lot Output (PLUTO) database to more accurately model the impact of limited access highways in New York City on asthma rates in the Bronx. PLUTO provides information on the various tax lots around NYC, including the number of residential units on a particular tax lot. For example, a tax lot with 10 units would thus have roughly double the number of occupants as a tax lot with 5 residential units. The fraction of the total population of the census tract residing on that particular tax lot would be determined by the number of residential units on that particular tax lot divided by the total number of residential units in the entire tract.

Assuming the two tax lots mentioned above were the only buildings in a particular census tract, the total number of residential units in the tract would be 15, with the first tax lot of 10 residential units having 2/3rds of the population and the second tax lot of 5 residential units having 1/3rd of the population. In the case of four year olds, if the population of four year olds was estimated to be 9, the first tax lot (of 10 residential units) would be estimated to have six four year olds while the second tax lot (with 5 units) would be assumed to have three four year olds. With this information, it would be possible to more accurately model the number of four year olds within the census tract instead of just assuming an even distribution across the entire tract where there could be large areas of non-residential space.

Accounting for public transit options that allow distant areas to be easily accessed in a reasonable amount of time would help provide a better estimate of the capacity available to a child at any given location. Incorporating travel by public transit would increase the number of hexagons to which a family could travel and increase the complexity of the calculation but more accurately estimate the real-world choices of parents.

Modeling the predominant direction of that flow would also help create a more reliable estimate of access, as the flow of New Yorkers tend to flow towards the business districts in lower and midtown Manhattan, downtown Brooklyn, and Long Island City in Queens, and back to the outer boroughs in
the evening, rather than flowing in all available directions equally at all times. Census journey to work data available through the Census Transportation Planning Products (CTPP) Program[^6] could help refine the analysis and give a better estimate of where parents would like to take their children.

Beyond these refinements to the methodology, having access to the data available in public records, such as birth, immunization, and public assistance records, could greatly improve the modeling techniques we employ, producing a much more accurate map of where four year olds live in New York City. As this is legally protected data, such work would need to be carried out under procedures outlined in federal, state, and local law, but could yield a highly accurate map of the four year old population in New York City.

6 Conclusion

We’ve outlined an approach that uses open data and open-source technology to reasonably estimate the distribution of four year olds in New York City using Census data and the available capacity in publicly funded pre-K programs using data from the New York City Department of Education. This approach gives decision makers the ability assess need in a quantitative way. Using additional demographic information, it becomes relatively easy to identify where that need exists, whether in low income areas, such as Sunset Park or higher income areas like the Upper East Side of Manhattan. Given limited resources, potential sites could be evaluated based on the likely impact on the local community given a set of criteria for evaluating the relative impact of one site over another, with the goal of providing capacity to those it’s likely to benefit most.

Our work is preliminary, with a number of methodological enhancements and improvements in data sources noted, but there are implications for how demand for school seats is modeled for all grades, providing a more sophisticated, data-driven approach to optimizing capacity to meet demand. We hope to make an interactive tool based on the model we’ve built in order to allow policymakers and government employees to adjust plans for new capacity in an easy and intuitive way.

The City of New York has embarked on an ambitious expansion of the pre-K program, the success of which will rely greatly on how well available technology is leveraged to achieve program success. To make the system truly universal, a seat should be available for each four year old at or near where they live. The realities of public finance and limitations of infrastructure and resources often mitigate against providing every public service at once and at the level expected. In that case, decision makers must target resources where they are needed most to alleviate the most critical need.

[^6]: Accessible at [http://ctpp.transportation.org/Pages/default.aspx](http://ctpp.transportation.org/Pages/default.aspx)
7 Acknowledgements

The authors would like to acknowledge Dr. Sharad Goel, Dr. Manuela Veloso, and Dr. Theodoros Damoulas for their assistance and stimulating conversations regarding this work, as well as the support of Dr. Constantine Kontokosta who supervised this research as part of the Urban Science Intensive I course at the Center for Urban Science and Progress at New York University.

References


UbiActive: A Smartphone-Based Tool for Trip Detection and Travel-Related Physical Activity Assessment

Yingling Fan*
Hubert H. Humphrey School of Public Affairs
University of Minnesota
301 19th Avenue South, Minneapolis, MN 55455
(612) 626-2930
yingling@umn.edu
*Corresponding author

Qian Chen
Hubert H. Humphrey School of Public Affairs
University of Minnesota
301 19th Avenue South, Minneapolis, MN 55455
chen2291@umn.edu

Chen-Fu Liao
Department of Civil Engineering
University of Minnesota
500 Pillsbury Drive Southeast, Minneapolis, MN 55455
cliao@umn.edu

Frank Douma
Hubert H. Humphrey School of Public Affairs
University of Minnesota
301 19th Avenue South, Minneapolis, MN 55455
douma002@umn.edu

Word count: 6176 words, 3 figures, 2 tables. (Equivalent: 7426 words.)
ABSTRACT
Utilizing the computing, communication, and sensing capabilities of smartphones, we developed an Android phone application—named UbiActive—to monitor travel behavior, assess travel-related physical activity, and generate daily activity summaries for smartphone users. We tested the application in the lab setting on three types of Android phones including HTC Magic, HTC MyTouch 4G, and SamSung Nexus S, as well as among a small group of real smartphone users who used a wide range of Android phones. Results from the tests confirmed UbiActive’s capability in collecting real-time data on travel behavior and physical activity as well as its capability in assessing and reporting travel-related physical activity. The tests show the potential to employ smartphone technology in increasing participants’ awareness of their travel behavior and travel-related physical activity. The tests also revealed rooms for improvement when it came to UbiActive’s trip detection function and its impact on batter consumption. All of these findings provide useful insights into future improvement of UbiActive and other smartphone-based applications.
INTRODUCTION

Regular physical activity reduces the risk of a number of diseases including coronary heart disease, diabetes, colon cancer, osteoarthritis, and osteoporosis (1). Yet, more than 80 percent of adults and youth in the U.S. do not meet the recommended guidelines for both aerobic and muscle-strengthening physical activity set by the U.S. Department of Health and Human Services (2). Public health professionals have increasingly looked beyond recreational pursuits for additional sources of physical activity, and emphasized a lifestyle approach to increasing activity that includes common behaviors such as brisk walking, climbing stairs, and doing house and yard work (3).

Among these activities, walking and bicycling for daily transportation have been considered as one of the most important sources of regular physical activity and present major opportunities for improving health of both children and adults (1). According to recent national travel behavior surveys, the U.S. has the lowest percentage of walking and bicycling trips (about 10%) and the highest driving mode share (about 85%) among wealthy countries (4, 5). And according to the 2010 American Time Use Survey, the U.S. population on average spent 66 minutes per person per day in cars (53 minutes as drivers and 13 minutes as passengers), and only 4 minutes per person per day on walking and biking trips. Time spent in cars is three times as much as thirty years ago. Given these unhealthy trends in U.S. travel behavior, it is especially important to understand the health-related impacts of one’s mode choice and more specifically the physical activity benefits associated with daily travel.

In this paper, we introduce an innovative, smartphone-based data collection tool that is capable of recording people’s daily travel behavior and more importantly assessing the amount of physical activity associated with their daily travel. The tool, named UbiActive, took advantage of smartphone’s portability as well as the computing, communication, and sensing capabilities of smartphones. Central to the UbiActive development was the utilization of various built-in smartphone sensors (e.g., GPS and accelerometer) for detecting trips and physical activity as well as a triggering mechanism that allows participants to self-report their trip characteristics immediately after the completion of each trip. After the development of UbiActive, we evaluated the application’s performance in the lab setting and among a small group of real smartphone users recruited from the University of Minnesota campus.

The paper is organized as follows: Section 2 summarizes the existing tools for monitoring travel behavior and physical activity. Section 3 provides an overview of the UbiActive application and three specific programs included in the application. Section 4 describes results from pilot testing of the application in the lab setting and evaluation experiments we conducted with real smartphone users. Section 5 offers concluding remarks.

EXISTING TOOLS FOR MONITORING TRAVEL BEHAVIOR AND PHYSICAL ACTIVITY

A variety of tools are commercially available to objectively detect and assess physical activity. The commonly used ones include pedometers and accelerometers. Pedometers are simple, inexpensive ($10-$50 per unit), and low-tech. Unlike accelerometers, pedometers are not designed to capture pattern, intensity, or type of physical activity. They do, however, detect steps taken with acceptable accuracy (6-8). Accelerometers are more expensive (the newer ones can cost as much as $500). Accelerometers also use much more sophisticated technology, many of
which possess a timing mechanism and a memory capacity that is capable of recording movement parameters (often expressed as activity counts) over researcher-determined units of time. The recorded movement parameters by accelerometers are often meaningless by themselves, and it is necessary to use cut points developed in laboratory studies to translate the parameters into estimates of activity duration in specific intensity categories. Technical expertise and additional hardware and software are required to calibrate, input, distill, and analyze data records directly downloaded from accelerometers.

Consequently, accelerometers are preferred tools for academic researchers to assess levels of physical activity among study participants. Pedometers are preferred tools for the general public to self-monitor their levels of physical activity. Indeed, pedometers have recently become a motivation tool for people wanting to increase their physical activity. Various websites exist to allow people to track their progress, including Fitbug (www.fitbug.com), Fitbit (www.fitbit.com), Philips Directlife (www.directlife.com), Mytrak (www.mytrak.com), Bodybugg (www.bodybugg.com), and Zamzee (www.zamzee.com).

More recently, interest has been rising in using modern mobile phones, especially smartphones with built-in accelerometer or pedometer, to collect physical activity data. This is largely due to the great advantage in data collection associated with mobile phones’ portability. Tracking physical activity using a device the user already carries/wears instead of an extra device is much more achievable for daily tracking as it can largely avoid missing data issues due to a participant not wearing the extra device. A number of interesting smartphone-based applications have been introduced in academia (9-12). For example, Bielik, P., et al. developed a system named Move2Play which serves to track activity and provides personalized recommendation and advice (12). It is worth noting that Move2Play focuses on walking activities that are monitored by pedometer embedded in smartphones (12). Also focusing on walking, Sohn et al. developed a software for GSM smart phones which is able to approximately count the user’s daily steps via the rate of change in cell tower observations (10). Besides walking, other types of physical activity have been studied as well. Soria-Morillo, L., et al. developed a system integrating mobile program to monitor “every movement practiced by the user and classify it in different activities such as walking, running, and jumping” (11). Shakra used the movement of mobile phones to track user’s active status and states of stationary, walking, and driving (9).

Interest has also been rising in using mobile phones to collect travel behavior data, especially after GPS tracking technology became widely available in commercial smartphones. Nonetheless, the adoption of this new technology is still slow. Existing smartphone research in the transportation field typically focuses on collecting travel behavior data without assessing travel-related physical activity (13-16), or focuses on providing information on available transportation services such as parking, ridesharing, and public transit (17-19). For example, trackit, a mobile application developed by researchers at the University of South Florida, was designed to collect highly accurate position data, transport this information to a server, and interact with users by allowing manual entry of survey data (20). Move, a smartphone application developed by research team at Ghent University, utilizes GPS and other information such as the “current and neighboring cell towers, the Wi-Fi stations in sight and their signal strengths” for location positioning, as well as built-in accelerometer to distinguish walking, biking and automobile modes (14). By communicating with cell towers and Wi-Fi stations, Move has the advantage of allowing tracking in both indoor and outdoor environments. However, its use of the built-in accelerometer is mainly for travel mode detection rather than physical activity.
assessment. As such, Move does not generate information on the amount of physical activity associated with daily travel behavior.

To the research team’s knowledge, none of the exiting smartphone-based applications were specifically targeted to assess travel-related physical activity except the Quantified Traveler application developed by a group of researchers at the University of California. However, similar to Move, the Quantified Traveler application did not use accelerometer outputs for physical activity assessment but travel mode detection. The application generated the travel-related physical activity measure using a calories calculator which adjusts calories burned by walking/biking speed. Driving, bus, and train trips were all assumed to burn zero calories. As such, the physical activity assessment in the Quantified Traveler application was simplistic and failed to capitalize on the sophisticated accelerometer technology that comes with smartphones for more robust physical activity assessment.

**UBIACTIVE SYSTEM OVERVIEW**

We used the Android operating system as the platform for developing UbiActive. Unlike Apple’s iPhones and Microsoft’s Palm series, Android phones offer an open development platform that gives programmers full access to the built-in hardware sensors including orientation, GPS, accelerometer, light, magnetic field, and temperature sensors. Our UbiActive application included three local programs running on the smartphones: a sensing and trip detection program, an after-trip survey program, and an activity summary program. The functions of each of the programs are summarized in Figure 1 and the details on the functions are described in the text below.

![FIGURE 1 UbiActive System Architecture Diagram](image-url)
**Program I: Sensing and Trip Detection**

This program was designed to automatically start and run in the background after each installation of the UbiActive application. It records time-specific acceleration, orientation, and location data non-stop from accelerometer, magnetic sensor, and GPS receiver at predetermined time intervals.

The built-in accelerometer in smartphones generates outputs in the unit of m/s². In UbiActive, the accelerometer was programmed to generate outputs every second (i.e., sampling frequency was configured to 1 Hz). Depending upon the Android version on the phone, the outputs were either linear acceleration measurements in x, y, and z directions with gravity force subtracted or acceleration measurements including the gravity force. In general, phones with Android version 2.3 or later provide linear acceleration outputs while the earlier versions do not. For phones with earlier Android versions, outputs from the magnetic sensor (sampling frequency also configured to 1 Hz) were used to subtract the gravity force from accelerometer outputs. The gravity-subtract accelerometer outputs were then used in Program III - Activity Summary (described below) to measure physical activity intensity, duration, and the associated energy expenditures.

The built-in GPS receiver in smartphones generates outputs in x and y coordinates. In UbiActive, the GPS receiver was programmed to generate outputs every 30 seconds. The generated time-specific location outputs were used to detect the beginning and completion of trips. Note that UbiActive was designed to record information only for trips longer than 10 minutes to avoid extremely short trips such as movements within the same building (e.g., walking to the cafeteria or restroom) as these short trips have limited relevance to daily travel routines. More specifically, to determine the occurrence of a valid trip (i.e., trip with duration longer than 10 minutes), we applied two different counters to GPS outputs: Counter A was for judging the start of a trip and Counter B was for determining the end of a trip.

With the GPS receiver updating locations every 30 seconds, it would report a “location change” event if it detected movement larger than 30 meters after 30 seconds. This event would be reported to Counter A. At that time, Counter A, whose default value was set to 0, would automatically add one. We used the 30-meter threshold for two reasons. First, according to the average walking speed of 1.4m/s (23, 24), the average 30-second movement by walking is around 42 meters. Therefore, choosing 30 meters per 30 second as a threshold ensures detection of movements by most travel modes. Second, GPS receivers usually have positioning errors, therefore, it is possible that the user is stationary, but GPS error leads to incorrect records of location change. Our testing experience showed that a 30-meter window helped to avoid incorrect records of location change. With this 30-meter threshold, a valid trip would be detected when Counter A reaches 20 counts, indicating there was a 10-minute continuous movement.

Once Counter A indicates the occurrence of a trip longer than 10 minutes, Counter B would start working to determine when the trip would end. Every 30 seconds, if no “location change” is updated, Counter B, whose default value was set 0, would automatically add 1. When Counter B reaches 10 counts, meaning there was no significant movement for 5 minutes, the trip would be considered as complete and Counter A would be reset to 0. Before Counter B detects the end of a trip (i.e. Counter B reaches 10 counts), any events of “location change” would lead to Counter B being reset to 0. In other words, we inferred the end of a valid trip as long as a stationary period of 5 minutes was observed. This 5-minute threshold is same as what Welbourne et.al. (2005) used in their research for determining the completion of a trip (25). Although some research used 10 minutes as the threshold (26), there has been concerns that a longer threshold...
would have higher chances to interrupt people’s after-trip activity. For example, if we trigger the survey 10 minutes after a participant finishes his/her trip, the participant is very likely already in the middle of another activity, like taking class or working, thus may not be available to answer the survey.

**Program II: After-Trip Survey**

The second program of UbiActive collected additional, self-reported trip attributes from participants through triggering survey actions on the smartphone immediately after Program I detected a trip. Each after-trip survey action was programmed to occur with beeper and vibration alerts.

The survey first asked participant to confirm whether he/she had completed a trip. It is possible that the detected movement is not a trip from the transportation standpoint, for example, participants might be jogging. An answer option of “it was not a valid trip” was provided for such cases. If the participant indicates no valid trip was made, the survey action would end. If the participant indicates the trip has not finished yet, the survey action would end as well. If the participant confirms completion of a valid trip, the participant would be further asked whether he/she is available to complete a survey. If yes, a series of questions would follow, as shown in Figure 2. If not, the participant would be asked “when do you prefer to answer the survey?” and then UbiActive re-triggers the survey at the time the participant chooses.

UbiActive allowed five minutes for the participant to respond each survey prompt. If a prompt is missed, UbiActive would send a second prompt to the participant 5 minutes later. If the participant misses the second prompt, the survey would be recorded as missing. The 5-minute waiting time threshold was selected based upon the existing literature (27, 28). In case UbiActive missed a valid trip and failed to automatically trigger an after-trip survey, UbiActive was designed to allow participants to self-trigger the survey by simply clicking an icon displayed on the home screen of their phones. As shown in Figure 2, each after-trip survey included two sets of questions. The first set (Questions 0-4) asked about basic trip information including start and end time, trip purpose, all travel modes used in order, accompany and secondary activities. The second set (Questions 5-8) asked about travel psychological experience/wellbeing.
FIGURE 2 Screenshots of the Interface of After-Trip Survey
Program III: Activity Summary

The third program in UbiActive was designed to summarize the real-time data collected in Programs I and II, and to provide daily activity summary reports to phone users. Since all the sensor-based logging data as well as the triggered survey data could be automatically stored to the local SQLite database running on the android phone (29), data processing for daily report was conducted locally on the android phone. As shown in Figure 3, the “Daily Report” summarized the following information: (a) total duration of physical activity in the past day; (b) total duration of physical activity related to travel; and (c) calories burned by travel and non-travel physical activity. It was scheduled that participants would receive daily reports at 10pm every day.

FIGURE 3 Screenshot Showing a Daily Physical Activity Report Example

This particular program involved converting linear acceleration outputs to energy expenditure measures. To do so, we first calculated activity count based upon linear acceleration outputs using the formula proposed by Bouten et al. (30).

\[
Activity\ Count = \int_{t_0}^{t_f} |a_x|\,dt + \int_{t_0}^{t_f} |a_y|\,dt + \int_{t_0}^{t_f} |a_z|\,dt
\]  
(Equation 1)

Where, \(a_x\), \(a_y\), \(a_z\) are the linear acceleration measurements in x, y, and z directions.

Then, to convert activity count data into energy expenditures, we calibrated our activity count data with energy expenditure data obtained from the commercially available RT3 accelerometer outputs. RT3 is a tri-axial accelerometer commercially available from StayHealthy, Inc. RT3 has been used by many researchers for measuring physical activity (31, 32). It provides physical activity data in the unit of both activity count and caloric expenditure. As for the calibration, to ensure that activity counts generated by smartphones and caloric energy expenditure outputs generated by RT3 measured the same activity, research assistants were asked to wear a RT3 unit and a Smartphone side by side on their right hip when conducting physical activity. The comparable RT3 energy expenditure outputs
and smartphone accelerometer outputs were then compared and used to calculate the RT3 scaling factor as shown below.

\[
RT3 \text{ Scaling Factor (kcal/count)} = \frac{RT3 \text{ Energy Expenditure (kcal/min)}}{\text{Smartphone Counts (count/min)}} \quad (\text{Equation 2})
\]

Derived from field experiments, the scaling factor was then used to convert activity count data from smartphones into energy expenditure data in calories (kcal) per min. In the current version of UbiActive, we calculated the cumulative energy expenditure per person per day using this method. We further divided the daily energy expenditure into travel-related and non-travel related expenditures. Energy expenditures associated with activities that had a corresponding travel speed higher than 1.5 m/s were considered to be energy expenditures associated with travel-related physical activity. Although the average walking speed is 1.4 m/s (23, 24), we chose a slightly higher threshold given that all of our participants are young adults.

**PILOT TESTING AND EVALUATION**

We tested UbiActive in the lab setting on three types of Android phones including HTC Magic, HTC MyTouch 4G, and Samsung Nexus S, as well as among a small group of real smartphone users who used a wide range of Android phones. All participants were recruited from the University of Minnesota campus using the convenience sampling technique in October 2011. The recruitment method included posting printed flyers around campus and sending emails to departmental listservs. An online screening survey was used to ensure that all participants had no health or disability problems preventing them from conducting active travel. Eligible participants were invited to a one-on-one orientation meeting which provided information on the general scope of this study, their responsibilities as a participant, confidentiality and voluntary nature of this research, as well as the amount of compensation they would receive. At the end of each orientation meeting, the researchers installed UbiActive on the participant’s phone with his/her consent. In addition, each participant was provided a paper-version travel diary booklet and was asked to fill in the diary at the end of the day. The purpose of this additional pen-and-paper instrument was to provide the respondent experiences in reporting travel data via both smartphones and the more traditional diary method so that they could make a comparison.

A total of 23 Android phone users were recruited in October, 2011 and 17 of them completed a 3-week trial of UbiActive in November, 2011. Table 1 below shows characteristics of the participants. The majority of the participants were male (71%), white (53%), full-time students (90%) between 19 and 30 years old. Vehicle access was not universal among the participants with 59% of them did not own a car. Almost everybody except one commuted to school or work by modes other than car. Yet, when it came to personal trips, 53% of the participants used cars.
In the following text, we describe major findings from our tests in the lab setting as well as from the three-week trials conducted by the 17 real Android phone users. The trials by real phone users generated two types of data:

a. raw sensing and self-reported data directly downloaded from participants’ phones at the end of their three-week participation, and

b. data from a web-based Exit Survey that participants were invited to fill out after their three-week participation to provide feedback on their general trial experience, perceived strength and weakness of UbiActive, and perceived difficulties in complying with the participation requirements.

### TABLE 1 Characteristics of participants in the three-week UbiActive trial (N=17)

<table>
<thead>
<tr>
<th>Variable</th>
<th>N (%)</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>5 (29%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>12 (71%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>9 (53%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>1 (6%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asian</td>
<td>5 (29%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>2 (12%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>23.35</td>
<td>2.95</td>
<td>19</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Occupation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full-time undergraduate student</td>
<td>7 (41%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full-time graduate student</td>
<td>8 (49%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part-time undergraduate student</td>
<td>1 (6%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alumni</td>
<td>1 (6%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smartphone brand</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HTC</td>
<td>4 (24%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LG</td>
<td>1 (6%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motorola</td>
<td>6 (35%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Samsung</td>
<td>6 (35%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vehicle ownership</td>
<td>7 (41%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main mode for work/school commute last week</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walking</td>
<td>8 (47%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bicycle</td>
<td>3 (18%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car</td>
<td>1 (6%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bus</td>
<td>5 (29%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Main mode for personal trips last week</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Walking</td>
<td>5 (29%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bicycle</td>
<td>1 (6%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Car</td>
<td>9 (53%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bus</td>
<td>2 (12%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Phone Performance

Tests in the lab setting were done on HTC Magic, HTC MyTouch 4G, and Samsung Nexus S phones. Results suggested a minimal impact of UbiActive on network usage with a consumption of data size less than 1KB per day. This consumption was largely due to the transfer of activity summary from smartphone to the server so that researchers could monitor for non-compliance issues and abnormal data patterns. As for memory requirement, UbiActive collected about 7Mb of raw sensor data and statistics per day. For a 3-week study period, the application needed about 150 Mb storage space which is sizable but easy to accommodate with most commercial phones on the market. The impact of UbiActive on battery life was significant: with the application running in the background continuously, the
battery lives of our testing phones were considerably shortened to about 12-15 hours without additional voice/text/data usage.

Participants in the three-week trials had no concerns about UbiActive’s impact on network and memory usage but its impact on phone battery life. Feedback from the web-based exit survey showed that the amount of battery consumed by UbiActive varied by phone. UbiActive was reported to shorten original phone battery life for a range between 25% and 83%. The newly released Android phones, such as Samsung Galaxy II, appeared to have longer battery life and were less influenced by UbiActive’s battery consumption than phones released earlier.

Sensing and Trip Detection Ability

In the lab setting, we conducted experiments and checked the quality of sensory outputs from HTC Magic, HTC MyTouch 4G, and SamSung Nexus S phones. We compared outputs from these phones’ built-in accelerometer with outputs from the commercially available RT3 accelerometer. Both HTC Magic and SamSung Nexus S phones provided good acceleration outputs. However, the HTC MyTouch phone, running on Android OS ver.2.2.1, would deactivate the internal sensory system if the touch screen is dimmed. This feature prohibited us from continuously collecting accelerometer and GPS data. To avoid this problem, we later added an eligibility criterion and asked only people with Android OS version earlier than 2.0 or later than 2.3 to participate in the three-week UbiActive trials.

We also conducted experiments to test UbiActive’s trip detection function. Two research assistants were asked to wear a smartphone (one with HTC Magic and another with Samsung Nexus S) with UbiActive application installed for 24 hours, and keep a diary to record the start and end time of each trip. Post-experiment data analysis showed that errors in trip detection were mainly caused by the GPS signal strength and multipath noises from the environment. The amount of errors depended heavily on the sampling frequency of the GPS receiver. Shorter sampling frequency say per second often generated false reports on trip completion while longer sampling frequency say per minute often led to misses of trip completion. After a series of experiments, we found that GPS position updates with a 30-second interval tended to generate the most accurate results.

Trials by real phone users suggested room for improvement when it came to UbiActive’s ability in trip detection. When asked “how satisfied are you with UbiActive’s ability of trip detection?”, only one (6%) participant reported “very satisfied”, five (29%) participants reported “satisfied”, five (29%) “somewhat satisfied”, and six (36%) reported “unsatisfied”. Participants’ feedback also showed that UbiActive’s ability in trip detection varied significantly by the type of phones and destinations people visit. For example, when asked “out of all the longer-than-10-minute trips you made during the three weeks of participation, for what percentage of these trips you were able to receive automatically triggered after-trip surveys?”, the reported auto-triggering rate of phone-based survey ranged from 0%-90%. The median of the reported auto-triggering rate was 50% with the 1st quartile at 35% and the 3rd quartile at 65%. The low auto-triggering rates were mainly reported by Motorola Droid and HTC Merge users, i.e. the earlier Android phones.
The variability in trip detection accuracy among different phones was not surprising because UbiActive’s ability of trip detection was highly dependent upon instant location updating of GPS sensor, which was determined partly by the accuracy of the GPS sensor in the phone and partly by the GPS signal strength at destination. And participants confirmed the triggering rate of phone-based survey varied significantly by the environment of destination. When asked “did you feel that the accuracy of trip detection varies by the kind of trips you made?” 59% participants answered “Yes”, and some of them specified the differences: “when destinations are close rooms, it never triggers”; “Trips ending in buildings seemed to be less likely to trigger a trip detection”; “The apps was very limited in triggering when traveling to buildings”.

Nonetheless, raw sensory outputs directly downloaded from the participants’ phones showed good data quality. Table 4 illustrates daily trip information captured from a participant on November 13, 2011. It was clear that data from the GPS and accelerometer sensors provided a deep understanding of participants’ travel behavior and travel-related physical activity, with origins, destinations, travel routes shown in Part I of Table 2 and travel speed and acceleration shown in Part II of Table 2.
### TABLE 2 Trip Information of Participant No. 15 on November 3, 2011 – Part I (location data and self-reported trip attributes)

<table>
<thead>
<tr>
<th></th>
<th>Trip #1</th>
<th>Trip #2</th>
<th>Trip #3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Start Time</strong></td>
<td>10:00am</td>
<td>12:00pm</td>
<td>4:30pm</td>
</tr>
<tr>
<td><strong>End Time</strong></td>
<td>10:15am</td>
<td>12:30pm</td>
<td>5:15pm</td>
</tr>
<tr>
<td><strong>Trip Purpose</strong></td>
<td>School</td>
<td>Meal</td>
<td>Back Home</td>
</tr>
<tr>
<td><strong>Mode</strong></td>
<td>Walking – Bus - Walking</td>
<td>Walking</td>
<td>Walking – Bus - Walking</td>
</tr>
<tr>
<td><strong>Companionship</strong></td>
<td>Alone</td>
<td>Alone</td>
<td>Friends/Schoolmates/</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Neighbors/Acquaintances</td>
</tr>
<tr>
<td><strong>Secondary Activity</strong></td>
<td>Doing Nothing</td>
<td>Doing Nothing</td>
<td>Talking/Conversation/Making Phone Call</td>
</tr>
<tr>
<td><strong>Satisfaction of Trip</strong></td>
<td>★★★★☆</td>
<td>★★★★☆</td>
<td>★★★★☆</td>
</tr>
<tr>
<td><strong>Does this trip make you feel good?</strong></td>
<td>★★★★☆</td>
<td>★★★★☆</td>
<td>★★★★☆</td>
</tr>
<tr>
<td><strong>Do positive aspects outweigh the negative of the trip?</strong></td>
<td>★★★★☆</td>
<td>★★★★☆</td>
<td>★★★★☆</td>
</tr>
<tr>
<td><strong>In general, how happy were you during this trip?</strong></td>
<td>★★★★☆</td>
<td>★★★★☆</td>
<td>★★★★☆</td>
</tr>
</tbody>
</table>

Pro. NSF Workshop on Big Data and Urban Informatics, Nov. 2015  Page 632
### TABLE 2  Trip Information of Participant No.15 on November 3, 2011 – Part II (travel speed and acceleration data)

<table>
<thead>
<tr>
<th>Trip #1</th>
<th>Trip #2</th>
<th>Trip #3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Walking - Bus – Walking</strong></td>
<td><strong>Walking</strong></td>
<td><strong>Walking – Bus - Walking</strong></td>
</tr>
</tbody>
</table>

**Trip 1 - Speed**

```
Walking  Walking  Waiting  Bus  Walking
```

**Trip 2 - Speed**

```
Walking  Waiting  Waiting  Walking
```

**Trip 3 - Speed**

```
Walking  Waiting  Bus  Walking
4:30:24 PM  4:40:25 PM  4:50:27 PM  5:00:28 PM  5:16:32 PM
```

**Trip 1 - Acceleration**

```
Walking  Waiting  Bus  Walking
```

**Trip 2 - Acceleration**

```
Walking  Waiting  Waiting  Walking
```

**Trip 3 - Acceleration**

```
Walking  Waiting  Bus  Walking
4:30:24 PM  4:40:25 PM  4:50:27 PM  5:00:28 PM  5:16:32 PM
```
After-Trip Survey and Activity Summary

The after-trip survey function of UbiActive performed well in both the lab and participant trial settings. Both research assistants and study participants reported higher willingness to take phone-based after-trip surveys than use the pen-and-paper instrument to report trip information. This was largely due to the sensing ability, portability, and real-time data entry function of smartphones. Participants also reported a high level of satisfaction with question wording, design of answer options, ease of navigation, and length of the after-trip survey. Thirteen (76%) participants strongly agreed with the statement that “answer options in the after-trip survey are reasonable and easy to pick”. Ten (59%) participants strongly agreed with the statement that “questions in the after-trip survey are easy to understand; eleven (65%) participants strongly agreed that “the after-trip survey is easy to navigate”; and nine (53%) participants strongly agreed that “the number of questions in the after-trip survey is about right (not too many)”. The activity summary function of UbiActive also performed well in both the lab and participant trial settings. Out of the seven participants who were provided daily activity summary, 6 participants agreed that the information provided in the reports was about right and easy to understand; 5 participants agreed that the information provided in the reports was useful and interesting, and made sense at the same time.

Overall participation Experience and Compliance Issues

Participants had high satisfaction with their UbiActive Trials. Thirteen (76%) participants reported they were “satisfied” with the overall experience. When asked “did your participation in this study make you more aware of your travel behavior?”, two (12%) participants reported “completely”, six (24%) reported “a great deal”, seven (53%) reported “somewhat”, and two (12%) reported “not at all”. When asked “did your participation in this study make you more aware of your travel-related physical activity?”, two (12%) participants reported “completely”, four (24%) reported “a great deal”, nine (53%) reported “somewhat”, and only two (12%) reported “not at all”.

Participants also did well with compliance requirements, except the requirement to fill out the paper-version travel diary and the requirement to manually trigger after-trip surveys when the phone fails to detect a trip. Only three (18%) of the participants reported to always or almost always comply with the travel diary requirement and only four (24%) reported to always or almost always comply with manual triggering requirement. When asked how disruptive the participation requirements were to their everyday life, participants suggested the most disruptive requirements to be timely battery charging, filling paper-version travel diary and keeping GPS setting always on. Most participants did not find wearing phone in close contact with body, answering phone-based surveys, keeping vibration setting on, and reading daily activity summary to be disruptive to their everyday life.
CONCLUSIONS

Overall, the UbiActive application worked well. Participants found it convenient to use and it provided a wealth of data that, until recently, could only be recorded in a cumbersome paper and pencil format. However, there is room for improvement. While the smartphones recorded most trips accurately, some trips were missed, and participants did not find a paper diary to be a user-friendly tool for verification and back-up.

Similarly, the requirement that UbiActive continually run in the background drained the battery more quickly than participants were used to, creating additional burdens. Finally, the question of whether receiving relatively rapid feedback regarding the health impact of one’s travel mode choice influences that choice is unclear given the extremely small sample size. At best, this research illustrates the potential to employing advanced smartphone technology in providing information and increasing awareness about travel-related physical activity. Potential impact of the proposed UbiActive application on awareness is unclear and cannot be generalized.

We conclude that UbiActive clearly contributes to the growing field of tools than can improve the collection of travel behavior data, and that further research into improvements that address the shortcomings noted above is warranted. Larger scale tests of these tools among the public could yield valuable information regarding the potential for use beyond the academic research community, potentially providing a new level of health and travel behavior data to transportation planners, health care providers and commercial interests. More importantly, this research used a new technology which is attracting new users each day. With nearly 50% of the U.S. population owning smartphones (and this percentage is likely to increase rapidly), smartphone applications like UbiActive could potentially have a broad public health impact to promote overall physical activity and reduce obesity in the general population.

REFERENCES


13. Tooley, M.S., *Using Smartphones to Collect Bicycle Travel Data in Texas*, 2011, University Transportation Center for Mobility, Texas Transportation Institute: College Station, TX.


Learning from Outdoor Webcams: Surveillance of Physical Activity Across Environments

J. Aaron Hipp, PhD\textsuperscript{1,2,*}, Deepti Adlakha, MUD\textsuperscript{1}, Amy A. Eyler, PhD\textsuperscript{1,2}, Rebecca Gernes, MSW, MPH\textsuperscript{1}, Agata Kargol\textsuperscript{3}, Abigail H. Stylianou\textsuperscript{3}, Robert Pless, PhD\textsuperscript{3}

Washington University in St. Louis, Missouri: \textsuperscript{1}Brown School, \textsuperscript{2}Prevention Research Center, and \textsuperscript{3}Computer Science and Engineering

*Corresponding author information: J. Aaron Hipp, PhD
One Brookings Drive, Washington University in St. Louis, Campus Box 1196
St. Louis, Missouri 63130
[o] 314.935.3868  [f] 314.935.8511  [e] ahipp@wustl.edu
**Keywords:** webcams, physical activity, built environment, crowdsourcing, outdoor environments

**Abstract**
There are tens of thousands of publicly available webcams which constantly view the world and share those images. These cameras include traffic cams, campus cams, ski-resort cams, etc. The Archive of Many Outdoor Scenes (AMOS) is a project that aims to geo-locate, calibrate, annotate, archive and visualize these cameras to serve as an imaging resource for a wide variety of scientific applications. Here we report on a multi-disciplinary project to demonstrate and evaluate the potential for webcams to be re-purposed as a tool to evaluate patterns of population-level physical activity behavior in diverse urban built environments.

The AMOS dataset has archived over 560 million images of outdoor environments from 27,000 webcams since 2006. The primary goal is to employ the AMOS dataset and crowdsourcing to develop reliable and valid tools to improve physical activity assessment via online, outdoor webcam capture of global physical activity patterns and urban built environment characteristics. This goal will be accomplished by addressing two subsequent aims:

**Aim 1:** Develop and test reliability of using publicly available, outdoor webcams to enumerate built environment characteristics and physical activity patterns across thousands of global outdoor environments.

**Aim 2:** Develop and test reliability and validity of using crowdsourcing to enumerate built environment characteristics and physical activity patterns across thousands of global outdoor environments.

This project’s grand scale-up of capturing physical activity patterns and built environments is a methodological step forward in advancing a real-time, non-labor intensive assessment using webcams and crowdsourcing. The combined use of webcams capturing outdoor scenes every 30 minutes and crowdsources providing the labor of annotating the scene allows for accelerated public health surveillance related to physical activity in built environments. The ultimate goal of this public health and computer vision collaboration is to develop machine learning algorithms that will automatically identify and calculate physical activity patterns.
INTRODUCTION

Kevin Lynch’s 1960 book, ‘The Image of the City’, was one of the first to emphasize the importance of social scientists and design professionals in signifying ways that urban design and built environment can be quantitatively measured and improved (Lynch, 1960). It led to enormous efforts to investigate the structure and function of cities, to characterize perception of neighborhoods (J. Jacobs, 1961; Xu, Weinberger, & Chapelle, 2012), and promotion of social interactions (Milgram, Sabini, & Silver, 1992; Oldenburg, 1989). To date, large scale studies seeking to understand and quantify how specific features or changes in the built environment impact individuals, their behavior, and interactions, have required extensive in-the-field observation. However, they only provide a limited view of behaviors, their context, and how each may change as a function of the built environment. These studies are time intensive and expensive, deploying masses of graduate students to conduct interviews about people’s daily routines (Milgram et al., 1992) or requiring hand-coding of thousands of hours of video (Whyte, 1980) to characterize a few city plazas and parks. Even current state-of-the-art technology to investigate associations between behavior and the urban built environment uses multiple expensive devices at the individual level (GPS and accelerometer) and connects this data to Geographic Information System (GIS) layers known to often be unreliable (James et al., 2014; Kerr, Duncan, & Schipperijn, 2011; Schipperijn et al., 2014).

A key population behavior of interest to our team is physical activity (Adlakha, Budd, Gernes, Sequeira, & Hipp, 2014; Eyler, Brownson, Schmid, & Pratt, 2010; Hipp, Adlakha, Eyler, Chang, & Pless, 2013). Physical activity plays a role in numerous health outcomes including obesity, diabetes, heart disease, and cancer (Office of the Surgeon General, 2011). Over 30% of adults and 17% of children and adolescents in the US are obese (CDC, 2009), with lack of physical activity due to constraints in the built environment being an important influence (O. Ferdinand, Sen, Rahurkar, Engler, & Menachemi, 2012). Lack of safe places to walk and bicycle and lack of access to parks and open space can impact the frequency, duration, and quality of physical activity of residents in urban settings (Brownson, Hoehner, Day, Forsyth, & Sallis, 2009; Jackson, 2003; Jackson, Dannenberg, & Frumkin, 2013). Physical activity may be purposive such as a jog in a park, or incidental such as a ten minute walk from home to a public transit stop. In both purposive and incidental cases the designs of urban built environments influence
the decisions and experience of physical activity behaviors. As such, the US Guide to Community Preventive Services (Community Guide) currently recommends the following built environment interventions to increase physical activity behaviors and reduce obesity: (1) community and street-scale urban design and land use policies; (2) creation of, or enhanced access to places for physical activity; and (3) transportation policies and practices (CDC, 2011).

**Physical Activity Assessment.** Physical activity and built environment research has expanded during the past 20 years (Handy, Boarnet, Ewing, & Killingsworth, 2002; O. Ferdinand et al., 2012). The research has followed traditional patterns of growth beginning with ecological studies of association (Ewing, Meakins, Hamidi, & Nelson, 2003), then local validation of associations via retrospective surveys and researcher-present observation (Bedimo-Rung, Gustat, Tompkins, Rice, & Thomson, 2006; McKenzie & Cohen, 2006). For example, the System for Observing Physical Activity and Recreation in Communities (SOPARC) (McKenzie & Cohen, 2006) was developed to understand physical activity in context with the environment while being unobtrusive. SOPARC continues to be a popular method of assessing physical activity with pairs of researchers positioning themselves in numerous target areas to scan the environment for numbers participating in sedentary to vigorous physical activity (Baran et al., 2013; Cohen, Marsh, Williamson, Golinelli, & McKenzie, 2012; Reed, Price, Grost, & Mantinan, 2012). Presently, natural experiments related to physical activity patterns and built environments are growing in popularity (Cohen et al., 2012). These studies have been of great benefit to the field by informing public health and urban design. While there is now a substantial body of evidence to inform local interventions and policies (Ding & Gebel, 2012; Feng, Glass, Curriero, Stewart, & Schwartz, 2010; Kaczynski & Henderson, 2007; Renalds, Smith, & Hale, 2010; Saelens & Handy, 2008; Sandercock, Angus, & Barton, 2010), currently used methodologies and the use of small, local samples limit the external validity and dissemination of many results, interventions, and policies. There is a need for large-scale, evidence-informed evaluations of physical activity to increase external validity as evident in recent calls for more studies across a greater variety of environments (Cerin, Conway, Saelens, Frank, & Sallis, 2009; Dyck et al., 2012).

**Big Data Opportunities.** Big data and modern technology has opened up several opportunities to obtain new insights on cities and offer the potential for dramatically more efficient
measurement tools (Graham & Hipp, 2014; Hipp, 2013). The relative ease of capturing large sample data has led to amazing results that highlight how people move through cities based on check-ins (Naaman, 2011; Silva, Melo, Almeida, Salles, & Loureiro, 2012) or uploaded photos (Crandall, Backstrom, Huttenlocher, & Kleinberg, 2009). In addition, GIS, GPS, accelerometers, smart phone applications (apps), and person-point-of-view cameras are each being used in many studies, often in combination (Graham & Hipp, 2014; Hurvitz, Moudon, Kang, Saelens, & Duncan, 2014; Kerr et al., 2011). Apps that track running and walking routes are being investigated for where populations move and how parks and other built environment infrastructure may be associated with such movement (Adlakha et al., 2014; Hirsch et al., 2014).

Though these big data sources offer important contributions to the field of physical activity and built environment research, they are each dependent on individuals to upload data, allow access to data, and/or agree to wear multiple devices. This is the epitome of the quantified-self movement (Barrett, Humblet, Hiatt, & Adler, 2013). A complementary alternative big data source is the pervasive capture of urban environments by traffic cameras and other public, online webcams. This environmental-point-of-view imaging also captures human behavior and physical activity as persons traverse and use urban space.

The Archive of Many Outdoor Scenes (AMOS) has been archiving one image each half hour from most online, publicly available webcams for the last 8 years, creating an open and widely distributed research resource (Pless & Jacobs, 2006). AMOS began to collect images from these 27,000 webcams mapped in Figure 1 to understand the local effects of climate variations on plants. We have used these large collections of up-close, on the ground measurements to suggest corrections to standard satellite data products like NASA’s Moderate Resolution Imaging Spectroradiometer (MODIS) estimates of tree growing seasons (Ilushin, Richardson, Toomey, Pless, & Shapiro, 2013; N. Jacobs et al., 2009; Richardson, Friedl, Froliking, Pless, & Collaborators, 2011). This global network of existing cameras also captures images of public spaces – plazas, parks, street intersections, waterfronts – creating an archive of how public spaces have changed over time and what behaviors are being performed within these spaces.
With its archive of over 550 million captured images, AMOS not only represents 27,000 unique environments, but is capturing concurrent behaviors in and across the environments. Unique and of significance to public health surveillance, the online, publicly available webcams are non-biased in data collection, consistent and thorough (an image each half hour), and timely (images instantly added to the archive and available to the public). The AMOS project provides an opportunity to virtually annotate changes in the built environment and associated physical activity behaviors. This dataset can provide a census of physical activity patterns within captured environments during the past eight years and moving forward. Due to the size of the AMOS dataset, we have used crowdsourcing to help annotate the captured scenes.

**Use of crowdsourcing in public health research.** Crowdsourcing refers to and utilizes the masses, or crowds, of individuals using the Internet, social media, and social smartphone apps. The crowds participating in these websites and applications are the source of data or the source of needed labor (Kamel Boulos et al., 2011). Crowdsourcing data collection in public health is an emerging field, with examples including the collection of tweets and Google searches that detected an increase in influenza before the increase in subsequent influenza-related hospital visits (Ginsberg et al., 2009; Kamel Boulos et al., 2011). Another potential use of crowdsourcing is as the labor in evaluation or assessment of research hypotheses (Bohannon, 2011; Buhrmester, Kwang, & Gosling, 2011; Office of the Surgeon General, 2011). The present team was the first to publish on the use of crowdsourcing as physical activity annotators (Hipp et al., 2013). A crowdsourcing marketplace, i.e., Amazon Mechanical Turk, can be used to ask workers to complete Human Intelligence Tasks (HITs) such as drawing a box around each pedestrian in a captured image.

**Objectives of Current Work.** The primary goal of our ongoing collaboration is to use the AMOS dataset and crowdsourcing to develop reliable and valid tools to improve physical activity behavior assessment. This goal will be accomplished by addressing two subsequent aims:

Aim 1: Develop and test the reliability of using publicly-available, outdoor webcams to enumerate built environment characteristics and physical activity patterns across thousands of global outdoor environments.
Aim 2: Develop and test the reliability and validity of using crowdsourcing to enumerate built environment characteristics and physical activity patterns across thousands of global outdoor environments.

**DATA SOURCES**

**Archive of Many Outdoor Scenes (AMOS).** The publicly captured scenes of human behavior, physical activity, and urban built environments are all from the AMOS dataset. AMOS is a Washington University project which aims to capture and archive images from every publicly available, online, outdoor webcam (e.g., traffic cams, campus cams, ski-resort cams, etc. – See Figure 1). This dataset was developed primarily as a basis to research computer vision algorithms for geo-locating and calibrating cameras, and as a demonstration that webcams can be re-purposed as a complement to satellite imaging for large-scale climate measurement (N. Jacobs et al., 2009; N. Jacobs, Roman, & Pless, 2008). Images are digitally captured from each camera every 30 minutes and archived in a searchable dataset.

Our current work builds on the AMOS model system for working with, sharing, and crowdsourcing big data. Figure 2 shows a screenshot of a main data access page, showing (A) one image and (B) the time this specific image was captured. A yearly summary image, indexed by time of year on the x-axis, and time of day on the y-axis is shown in (C). This summarizes a year of images with each pixel as a representation of the image at that time of year and time of day. Pixels can also be represented using principal component analysis to quickly identify images that differ based on precipitation, snowfall, dusk, dawn, etc. This summary serves several purposes. First, it is a data availability visualization, where dark red highlights when the camera was down and did not capture images. Second, it highlights annual patterns such as the summer nights being shorter than winter nights. Third, data capture problems are often visible. Finally, this data visualization is “clickable” so that a user can see, by clicking, the image from a particular time of day and time of year.

Each camera also contains extensive metadata as outlined in the Figure 2: (D) Shows the geo-location of the camera; (E) Shows free form text tags that we and other groups use to keep track of and search for cameras with particular properties; (F) is a new feature added for this present
project that allows the tagging of specific images (instead of cameras), and \( G \) is a pointer to zip-files for data from this camera or a python script to allow selective downloading. When exact camera locations are known, the cameras may be geo-oriented and calibrated relative to global coordinates as shown in Figure 1.

**Amazon.com’s Mechanical Turk Crowdsourcing.** Virtual audits have emerged as a reliable method to process the growing volume of web-based data on the physical environment (Badland, Opit, Witten, Kearns, & Mavoa, 2010; Clarke, Ailshire, Melendez, Bader, & Morenoff, 2010; C. L. Odgers, A. Caspi, C. J. Bates, R. J. Sampson, & T. E. Moffitt, 2012). Research has also turned to virtual platforms as a way to recruit study participants and complete simple tasks (Hipp et al., 2013; Kamel Boulos et al., 2011). The Amazon.com Mechanical Turk (MTurk) website outsources Human Intelligence Tasks (HITs), or tasks that have not yet been automated by computers. Workers may browse available HITs and are paid for every HIT completed successfully (Buhrmester et al., 2011). MTurk workers are paid a minimum of US$0.01 per HIT, making them a far less expensive option than traditional research assistant annotators (Berinsky, Huber, & Lenz, 2012). MTurk was found as an effective method for survey participant recruitment, with more representative and valid results than the convenience sampling often used for social science research (Bohannon, 2011). MTurk has also been used for research task completion such as transcription and annotation. These have generally been small in scale and MTurk reliability for larger scale data analysis has not been established (Hipp et al., 2013). Within MTurk, our team has designed a unique web-form used with the MTurk HIT that allows amateur workers to annotate our images by demarcating each pedestrian, bicyclist, and vehicle per photograph.

**Trained Research Assistants.** Trained undergraduate and graduate Research Assistants from the computer science and public health departments at Washington University in St. Louis have annotated images for physical activity behaviors and built environment attributes. For both behaviors and environments, Research Assistants were provided with example captured scenes. Project Principal Investigators supervised the scene annotation process and provided real-time feedback on uncertain scenes. Difficult or exceptional scenes and images were presented to the
research group to ensure that all behaviors and environments were annotated in a consistent manner.

**METHODS**

**Annotating Physical Activity Behaviors.** We have used 12 traffic webcams located in Washington, DC, to determine initial feasibility of the physical activity behavior research agenda. AMOS has archived a photograph every thirty minutes from Washington, DC, Department of Transportation webcams. Since 2007, Washington, DC, has initiated multiple built environment improvements to increase physical activity behaviors, including a bicycle share program, miles of new bike lanes, and painted crosswalks. For example, a new bicycle lane was added in the middle of Pennsylvania Avenue in spring 2010, and AMOS has an archive of captured images every thirty minutes for the year prior to the installation of the bike lane, and a year following installation.

The MTurk website was used to crowdssource the image annotation. In a pilot project we uploaded each webcam photograph captured by AMOS at the intersection of Pennsylvania Avenue NW and 9th Street NW between 7am and 7pm the first work week of June 2009 and June 2010 to the MTurk webpage (Hipp et al., 2013). There we designed a HIT that allowed MTurk workers to annotate our images by marking each pedestrian, bicyclist, and vehicle in each captured scene. MTurk workers used their computer mouse to hover over the appropriate behavior, e.g., pedestrian activity, and left-click atop each individual pedestrian. Five unique MTurk workers completed this task for all three transportation behaviors per image. The numbers of each type of annotation were then downloaded to a spreadsheet and imported into SPSS.

In related ongoing work, we have used 12 different AMOS webcams that captured other built environment changes at intersections in Washington, DC, between 2007 and 2010. We have made improvements to our MTurk task by asking workers to use their cursors to draw polygons, or boxes, around the specified transportation behavior (walking, cycling, driving). Similar to the first HIT, we used each photograph between 7:00am and 7:00pm during the first week of June proceeding and following a built environment change. Finally, we posted to MTurk every
photograph from two of the above intersections between 6:00am and 9:00pm for 19 consecutive months (five months prior to a crosswalk being introduced to the intersections and 14 months post).

MTurk workers were paid US$0.01 or US$0.02 per scene to mark each pedestrian, cyclist, and vehicle in an image and took on average 71 seconds to complete each task. Each image was annotated five unique times. Two trained Research Assistants completed the same task, annotating each image twice. Training took place in two sessions. In the first session, Research Assistants received the same instructions as MTurk participants and completed a practice set of 100 images. In the second session, Research Assistants compared their practice results and discussed differences in analysis. Research Assistants completed the camera annotations in separate forms, and their results were averaged.

**Annotating Built Environments.** Selecting the appropriate built environment image tags was an iterative process. First, we selected two commonly used built environment audit tools to establish a list of potential built environment tags. These were the Environmental Assessment of Public Recreation Spaces (EAPRS) (Saelens et al., 2006) and the Irvine-Minnesota Inventory (Day, Boarnet, Alfonzo, & Forsyth, 2006). From an initial list of 73 built environment items that we believed could be annotated using captured images we narrowed the final list down to 21 built environment tags. Following the combination of similar terms, we further reduced the potential list of tags based on the inclusion criteria that the tag must be theoretically related to human behaviors.

To establish which of the 27,000 AMOS webcams are at an appropriate urban built environment scale, i.e., those with the potential of capturing physical activity, our team designed an interface that selects a set of camera IDs, and displays 25 cameras per screen. This internal HIT was created to populate a webpage with the 25 unique camera images. Below each image was a green checkmark and a red x-mark. If physical activity behaviors could be captured in the scene, the green checkmark was selected and this tag automatically added to a dataset of physical activity behavior cameras. This process was repeated with trained Research Assistants for reliability and resulted in a set of 1,906 cameras. In addition to the above inclusion criteria, selected cameras
must have captured scenes from at least 12 consecutive months. The final 21 built environment tags are presented in Table 1.

To tag each camera, Research Assistants were provided a one-page written and photographic example (from AMOS dataset) of each built environment tag. For example, a written description for a crosswalk was provided along with captured images of different styles of crosswalks from across the globe. A second internal HIT was created similar to the above that populated a webpage with 20 unique camera images, each marked with a green checkmark and a red x-mark. If the provided built environment tag (e.g., crosswalk) was present in the image then the green checkmark was selected and this tag automatically added to the camera annotation. If a Research Assistant was unsure they could click on the image to review other images captured by the same camera or could request the assistance of other Research Assistants of Principal Investigators to verify their selection. This process was completed for all 21 built environment tags across all 1,906 cameras in the AMOS physical activity dataset. To date, the built environment tags have only been annotated by trained Research Assistants. Reliability and validity of tags is a future step of this research agenda. This initial step provided the team a workable set of publicly available webcams to address our two study aims.

DATA ANALYSIS

Physical Activity Behaviors. In the pilot project we used t-tests and logistic regressions to analyze the difference in physical activity behaviors before and after the addition of the bike lane along Pennsylvania Avenue. T-tests were used for pedestrians and vehicles, where the data was along a continuous scale from 0-20 (20 being the most captured in any one scene). Logistic regression was used for the presence or absence of a cyclist in each image.

Reliability and Validity. Inter-rater reliability (IRR) and validity statistics (Pearson’s R, Inter-Class Correlations, and Cohen’s Kappa) were calculated within and between the five MTurk workers and between the two trained Research Assistants. The appropriate statistic was calculated for two, three, four, or five MTurk workers to determine the optimal number of workers necessary to capture a reliable and valid count of pedestrians, cyclists, and vehicles in a scene. Due to each scene being annotated by five unique MTurk workers we were able to test the
reliability of ten unique combinations of workers; that is, Worker 1 and Worker 2, Worker 1 and Worker 3, Worker 1 and Worker 4, etc. Similar combinations were used with three workers (ten unique combinations) and four workers (five unique combinations). Each combination was compared to the trained Research Assistants results to measure validity. For all tests we used Landis and Koch’s magnitudes of agreement: <0.19 (poor agreement), 0.20-0.39 (fair), 0.40-0.59 (moderate), 0.60-0.79 (substantial) and >0.80 (near perfect agreement) (Landis & Koch, 1977).

RESULTS

Pilot Project. Previously published results reveal that publicly available, online webcams are capable of capturing physical activity behavior and are capable of capturing changes in these behaviors pre and post built environment changes (Hipp et al., 2013). The odds of the traffic webcam at Pennsylvania Avenue NW and 9th Street NW capturing a cyclist present in the scene in 2010 increased 3.5 times, compared to 2009 (OR=3.57, p<0.001). The number of cyclists per scene increased four-fold between 2009 (mean=0.03; SD=0.20) and 2010 (0.14; 0.90; F=36.72, 1198; p=0.002). Both results are associated with the addition of the new bike lane. There was no associated increase in the number of pedestrians at the street intersection following the addition of the bike lane, as may be theoretically expected with a bicycle-related built environment change, not a pedestrian-related change.

Reliability Assessment. Next, we tested reliability and validity of using publicly available webcams and MTurk HITs to annotate captured scenes for physical activity and transportation behaviors. Reliability statistics varied across MTurk workers based on the number annotating each scene and the annotation task (pedestrians compared to cyclists).

For pedestrians (n = 720 images), pairs of MTurk workers had an agreement average and a Pearson’s R-score of 0.562 (range: 0.122 – 0.866). The Inter-Class Correlation (ICC) for three MTurk workers averaged 0.767 (0.330 – 0.944) and four workers averaged 0.814 (0.534 – 0.954). The average for all five workers across the 720 scenes was 0.848 (0.687 – 0.941). The ICCs for four and five workers represented near-perfect agreement. The pair of trained Research Assistants averaged a Pearson’s R-score 0.850 (0.781 – 0.925), also representing near perfect agreement.
The averages and ranges of annotator agreement for presence of cyclists in 2007 were as follows (Table 2): two workers (Cohen’s Kappa: 0.333; Range: 0.000 – 0.764), three workers (0.553; 0.000 – 0.897), four workers (0.607; 0.000 – 0.882), five workers (0.645; 0.000 – 0.874), and Research Assistants (0.329; 0.000 – 0.602). Annotator agreement with four and five MTurk workers showed substantial agreement. For the pilot project presented above, we used the average of five MTurk workers. When analyzing presence versus absence, majority ruled; if three or more of the five MTurk workers annotated a cyclist was present, then this scene received a 1. If two or fewer annotated a cyclist, the scene received a 0.

The averages and ranges for number of vehicles were as follows: two workers (0.354; 0.000 – 0.769), three workers (0.590; 0.208 – 0.830), four workers (0.653; 0.398 – 0.830), five workers (0.705; 0.592 – 0.837), and Research Assistants (0.885; 0.841 – 0.922). The reliability statistics for four and five MTurk workers again showed substantial rater/annotator agreement, and near perfect agreement between the two Research Assistants.

**Validity Assessment.** From reliability estimates, we concluded that using four MTurk workers was the most reliable and cost-efficient method. Next, validity statistics were calculated for four MTurk workers and two trained RAs. Validity statistics (Pearson’s R) for pedestrians (0.846 – 0.901) and vehicles (0.753 – 0.857) showed substantial to near perfect agreement. Validity (Cohen’s kappa) for cyclists (0.361 – 0.494) were in the fair-moderate agreement range.

**Built Environment Tags.** As provided in Table 1, our final list of built environment tags includes 21 unique items. The number of cameras with the tag present is also presented. ‘Buildings’ was found the most frequent, present and at a scale to capture human behavior across 1,245 webcams. ‘Bike racks’ was annotated the fewest times, only occurring in 27 scenes. Figure 4 shows an example map of where each of the cameras with the built environment tag of ‘open space’ is located.

**DISCUSSION**

---

Proc. NSF Workshop on Big Data and Urban Informatics, Nov. 2015  Page 650
The use of public, captured imagery to annotate built environments for public health research is an emerging field. To date the captured imagery has been static and only available via Google Streetview and Google Satellite imagery (Charreire et al., 2014; Edwards et al., 2013; Kelly, Wilson, Baker, Miller, & Schootman, 2012; Kelly et al., 2014; Candice L. Odgers, Avshalom Caspi, Christopher J. Bates, Robert J. Sampson, & Terrie E. Moffitt, 2012; Rundle, Bader, Richards, Neckerman, & Teitler, 2011; B. T. Taylor et al., 2011; J. R. Taylor & Lovell, 2012; Wilson & Kelly, 2011; Wilson et al., 2012). There have been no attempts to crowdsource this image annotation, nor combine annotation of built environments and images capturing physical activity behaviors. Using an eight-year archive of captured webcam images and crowdsources, we have demonstrated that improvements in urban built environments are associated with subsequent and significant increases in physical activity behaviors. Webcams are able to capture a variety of built environment attributes and our study shows webcams are a reliable and valid source of built environment information. As such, the emerging technology of publicly available webcams facilitates both consistent uptake and potentially timely dissemination of physical activity and built environment behaviors across a variety of outdoor environments. The AMOS webcams have the potential to serve as an important and cost-effective part of urban environment and public health surveillance to evaluate patterns and trends of population-level physical activity behavior in diverse built environments.

In addition to presenting a new way to study physical activity and the built environment, our findings contribute to novel research methodologies. The use of crowdsources (Amazon’s Mturk) proved to be a reliable, valid, inexpensive, and quick method for annotating street scenes captured by public, online webcams. While MTurk workers have previously been found to be a valid and reliable source of participant recruitment for experimental research, this is the first research agenda that has found MTurk to be a valid and reliable method for content analysis (Buhrmester et al., 2011, Berinsky et al., 2012, Hipp et al., 2013). Our results indicate taking the average annotation of four unique MTurk workers appears to be the optimal threshold. Our results also show that across each mode of transportation assessed, the average reliability score with four unique workers was 0.691, which is considered substantial agreement (Landis & Koch, 1977).
In addition to substantial agreement between the MTurk workers, the trained RAs yielded substantial agreement with vehicles, near perfect agreement with pedestrians, but only fair agreement with cyclists. The cyclists’ statistics were the least reliable, primarily due to the low number of images (only 10% of captured scenes) with a cyclist present. Similar to reliability statistics, validity was near perfect for pedestrians and vehicles, but only fair to moderate for cyclists. These results suggest MTurk workers are a quick, cheap annotation resource for commonly captured image artifacts. However, MTurk is not yet primed to capture rare events in captured scenes without additional instruction or limitations to the type of workers allowed to complete tasks.

Our current big data and urban informatics research agenda shows that publicly available, online webcams offer a reliable and valid source for measuring physical activity behavior in urban settings. Our findings lay the foundation for studying physical activity and built environment characteristics using the magnitude of available globally recorded images as measurements. The research agenda is innovative in: (1) its potential to characterize physical activity patterns over the time-scale of years, with orders of magnitude more measurements than would be feasible by standard methods, (2) the ability to use the increase in data to characterize complex interactions between physical activity patterns, seasons and weather, and (3) its capacity to be an ongoing, systematic public health surveillance system. In addition to increasing the capacity of physical activity research, the methodologies described here are of novel interest to computer vision researchers. Automating algorithms to detect and quantify behavioral transformations due to changes in urban policy and built infrastructure can transform aspects of this research as well.

These findings have several implications related to cost and timeliness for the use of MTurks in content analysis. The total cost for the MTurk analysis was $320.00 for 32,001 images, compared to $1,333.33 for a trained Research Assistant paid at US$10 per hour. This could be due to several cost-saving characteristics of crowdsourcing. MTurk workers are paid for each successful completion of a HIT, compared to hourly wages for a Research Assistant. This allows MTurk to pay multiple workers at a time for each HIT, at a cost substantially lower than the same number of trained Research Assistants. The higher speed and lower cost of crowdsourced analysis is especially suitable for annotating AMOS data, to which thousands of images are
added daily. Reliable and rapid image annotation using MTurks could allow for large-scale and
more robust analysis of results that would be too costly to complete with traditional analysis.
Thus far our team has looked at 12 cameras in one metro area. Future studies could increase the
number of cameras annotated for a specific area or time, compare results across metro regions, or
analyze environmental effects such as weather, season, and day of the week on mode of
transportation.

There are several ethical and human subjects concerns related to publicly available, online
webcams and the use of MTurk. With our initial research projects, we have received exempt
status for the use of both AMOS and MTurk. The webcams were exempt because we are not
collecting individual identifiable private information; this activity is not considered to meet
federal definitions under the jurisdiction of an Institutional Review Board and therefore falls
outside the purview of the Human Research Protection Office. AMOS is an archival dataset of
publicly available photos. The photos are being used for counts and annotation of physical
activity patterns and built environment attributes and are not concerned with individual or
identifiable information. To date, no camera has been identified that is at an angle and height so
as to distinguish an individual’s face. The use of publicly available webcams fits with the ‘Big
Sister’ approach to the use of cameras for human-centered design and social values (Stauffer &
Grimson, 2000). Related, recent research utilizing Google Street View and Google Earth images
have also been HRPO-exempt ("National Center for Safe Routes to School," 2010; Sadanand &
Corso, 2012; Saelens et al., 2006; Sequeira, Hipp, Adlakha, & Pless, 2013).

Finally, AMOS is quite literally “Seeing Cities Through Big Data” with applications for research
methods and urban informatics. With thoughtful psychometrics and application of this half-
billion image dataset, and growing, we believe pervasive webcams can assist researchers and
urban practitioners alike in better understanding how we use place and how the shape and
context of urban places influence our movement and behavior.
References


Ding, Ding, & Gebel, Klaus. (2012). Built environment, physical activity, and obesity: What have we learned from reviewing the literature? *Health & Place, 18*(1), 100-105. doi: 10.1016/j.healthplace.2011.08.021


Naaman, Mor. (2011). Geographic Information from Georeferenced Social Media Data. SIGSPATIAL Special, 3(2), 54-61.


Oldenburg, Ray. (1989). The great good place: Cafés, coffee shops, community centers, beauty parlors, general stores, bars, hangouts, and how they get you through the day: Paragon House New York.


Sequeira, Sonia, Hipp, Aaron, Adlakha, Deepti, & Pless, Robert. (2013). Effectiveness of built environment interventions by season using web cameras *141st APHA Annual Meeting (November 2-November 6, 2013)*.


Figure 1. Map of cameras captured by the Archive of Many Outdoor Scenes (AMOS).
Figure 2. Screenshot of an AMOS data access page.
Figure 3. Reliability results for annotation of pedestrians in 720 webcam scenes.
Figure 3. Location of AMOS webcams tagged with ‘open space’.
Table 1. List of Built Environment tags used to annotate AMOS webcam images.

<table>
<thead>
<tr>
<th>No.</th>
<th>Built Environment Tag</th>
<th>Number of Cameras with Built Environment Tag Present</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Open space</td>
<td>769</td>
</tr>
<tr>
<td>2.</td>
<td>Sidewalk</td>
<td>825</td>
</tr>
<tr>
<td>3.</td>
<td>Plaza/ Square</td>
<td>174</td>
</tr>
<tr>
<td>4.</td>
<td>Residential/ Homes</td>
<td>97</td>
</tr>
<tr>
<td>5.</td>
<td>Trees</td>
<td>1,058</td>
</tr>
<tr>
<td>6.</td>
<td>Buildings</td>
<td>1,245</td>
</tr>
<tr>
<td>7.</td>
<td>Street, intersection</td>
<td>621</td>
</tr>
<tr>
<td>8.</td>
<td>Bike lane</td>
<td>71</td>
</tr>
<tr>
<td>9.</td>
<td>Athletic fields</td>
<td>60</td>
</tr>
<tr>
<td>10.</td>
<td>Speed control</td>
<td>185</td>
</tr>
<tr>
<td>11.</td>
<td>Trail path</td>
<td>154</td>
</tr>
<tr>
<td>12.</td>
<td>Street, road</td>
<td>1,029</td>
</tr>
<tr>
<td>13.</td>
<td>Signage</td>
<td>59</td>
</tr>
<tr>
<td>14.</td>
<td>Commerce Retail</td>
<td>382</td>
</tr>
<tr>
<td>15.</td>
<td>Play features</td>
<td>42</td>
</tr>
<tr>
<td>16.</td>
<td>Sitting features</td>
<td>166</td>
</tr>
<tr>
<td>17.</td>
<td>Motor vehicles</td>
<td>1,048</td>
</tr>
<tr>
<td>18.</td>
<td>Crosswalk</td>
<td>576</td>
</tr>
<tr>
<td>19.</td>
<td>Bike racks</td>
<td>27</td>
</tr>
<tr>
<td>20.</td>
<td>Water</td>
<td>326</td>
</tr>
<tr>
<td>21.</td>
<td>Snow</td>
<td>169</td>
</tr>
</tbody>
</table>
Table 2: Inter-rater reliability coefficients for trained Research Assistants and Mechanical Turk workers.

<table>
<thead>
<tr>
<th>Camera ID</th>
<th>Counts</th>
<th>N</th>
<th>RAs Correlation</th>
<th>2 MTurkers Range</th>
<th>3 MTurkers Correlation Average</th>
<th>4 MTurkers ICC Range</th>
<th>5 MTurkers ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>919</td>
<td>Pedestrian</td>
<td>74</td>
<td>.779**</td>
<td>.122 - .769**</td>
<td>.344</td>
<td>.330*** - .825***</td>
<td>.687***</td>
</tr>
<tr>
<td></td>
<td>Bicyclist</td>
<td>74</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td>Motor Vehicle</td>
<td>74</td>
<td>.916**</td>
<td>.122 - .769**</td>
<td>.344</td>
<td>.330*** - .663***</td>
<td>.687***</td>
</tr>
<tr>
<td>920</td>
<td>Pedestrian</td>
<td>121</td>
<td>.925**</td>
<td>.382** - .800**</td>
<td>.604</td>
<td>.724*** - .881***</td>
<td>.819*** - .874***</td>
</tr>
<tr>
<td></td>
<td>Bicyclist</td>
<td>121</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
<td>.267* - .773***</td>
<td>.507*** - .737***</td>
</tr>
<tr>
<td></td>
<td>Motor Vehicle</td>
<td>121</td>
<td>.852**</td>
<td>.013 - .502**</td>
<td>.235</td>
<td>.208*** - .630***</td>
<td>.398*** - .614***</td>
</tr>
<tr>
<td>929</td>
<td>Pedestrian</td>
<td>120</td>
<td>.822**</td>
<td>.342** - .812**</td>
<td>.577</td>
<td>.706*** - .897***</td>
<td>.812*** - .882***</td>
</tr>
<tr>
<td></td>
<td>Bicyclist</td>
<td>120</td>
<td>.524***</td>
<td>.425** - .635***</td>
<td>.566</td>
<td>.706*** - .897***</td>
<td>.882*** - .857***</td>
</tr>
<tr>
<td></td>
<td>Motor Vehicle</td>
<td>120</td>
<td>.902**</td>
<td>.257** - .702**</td>
<td>.508</td>
<td>.605*** - .838***</td>
<td>.730*** - .841***</td>
</tr>
<tr>
<td>930</td>
<td>Pedestrian</td>
<td>125</td>
<td>.900**</td>
<td>.659** - .860**</td>
<td>.784</td>
<td>.875*** - .944***</td>
<td>.917*** - .954***</td>
</tr>
<tr>
<td></td>
<td>Bicyclist</td>
<td>125</td>
<td>.547***</td>
<td>.317** - .663***</td>
<td>.426</td>
<td>.477*** - .685***</td>
<td>.578*** - .756***</td>
</tr>
<tr>
<td></td>
<td>Motor Vehicle</td>
<td>125</td>
<td>.878**</td>
<td>.177** - .577**</td>
<td>.340</td>
<td>.390*** - .730***</td>
<td>.575*** - .685***</td>
</tr>
<tr>
<td>942</td>
<td>Pedestrian</td>
<td>126</td>
<td>.781**</td>
<td>.355** - .534**</td>
<td>.450</td>
<td>.679*** - .749***</td>
<td>.765*** - .780***</td>
</tr>
<tr>
<td></td>
<td>Bicyclist</td>
<td>126</td>
<td>.602***</td>
<td>.281** - .764***</td>
<td>.552</td>
<td>.680*** - .868***</td>
<td>.783*** - .871***</td>
</tr>
<tr>
<td></td>
<td>Motor Vehicle</td>
<td>126</td>
<td>.841**</td>
<td>.099 - .411**</td>
<td>.269</td>
<td>.392*** - .617***</td>
<td>.544*** - .634***</td>
</tr>
<tr>
<td>996</td>
<td>Pedestrian</td>
<td>128</td>
<td>.893**</td>
<td>.526** - .709**</td>
<td>.615</td>
<td>.790*** - .856***</td>
<td>.851*** - .877***</td>
</tr>
<tr>
<td></td>
<td>Bicyclist</td>
<td>128</td>
<td>.301***</td>
<td>.169 - .659**</td>
<td>.317</td>
<td>.437*** - .788***</td>
<td>.645*** - .792***</td>
</tr>
<tr>
<td></td>
<td>Motor Vehicle</td>
<td>128</td>
<td>.922**</td>
<td>.291** - .682**</td>
<td>.381</td>
<td>.619*** - .742***</td>
<td>.690*** - .774***</td>
</tr>
</tbody>
</table>

Significance levels: *p < .05, **p < .01, ***p < .001
Abbreviations: RA: Research Assistant; ICC: Intraclass Correlation Coefficients
*Pearson correlations were used in this calculation
*Kappas were used in this calculation
*Items could not be calculated due to insufficient variance between raters
Extracting activity patterns from GPS track data

Sungsoon Hwang (DePaul University, Geography, shwang9@depaul.edu)
Christian Evans (Midwestern University, Physical Therapy, CEvans@midwestern.edu)
Timothy Hanke (Midwestern University, Physical Therapy, THANKE@midwestern.edu)

Abstract
This paper considers accuracy of trip detection algorithms in classifying GPS track points into significant trips and stops. The proposed method modifies DBSCAN, a density-based spatial clustering algorithm, to temporal criteria to detect stops as spatiotemporal clusters, and uses temporal filtering to smooth out any misclassified values. The proposed method was applied to twenty week data from seven subjects who agree to carry a GPS data logger for a week continuously. Percent correctly classified based on the proposed method is 93.8% with kappa index .87. Experimentation results indicate that a clustering algorithm combined with smoothing techniques provides a relatively reliable means to infer events from noisy GPS track data.

Keywords: GPS, trip detection, DBSCAN, spatial trajectories, big data

1. Introduction
This paper presents a method for extracting activity patterns from GPS track data. GPS track data refer to a collection of track points with coordinates recorded at a predefined interval. These data have been used to complement personal travel survey (Stopher et al., 2008), detect activity patterns of individuals (Eagle & Pentland, 2006), examine effects of transportation infrastructure on physical activity (Duncan et al., 2009), improve measurement of physical activity for health applications (Krenn et al. 2011), and understand role of environmental factors (e.g., pollution) in occurrence of diseases (e.g., cancer, asthma) (Richardson et al., 2013).

The utility of GPS tracking will only grow as greater amount of GPS data become readily available from location-aware devices. With increasing population in urban areas, the ability to track human movements in high spatiotemporal resolution offers much potential in urban informatics—“the study, design, and practice of urban experiences across different urban
contexts that are created by new opportunities of real-time, ubiquitous technology and the augmentation that mediates the physical and digital layers of people networks and urban infrastructures” (Foth et al, 2011). GPS data--being integrated with other data such as survey, census, and GIS data--can lead to a better understanding of “circulatory and nervous systems” of cities (Townsend, 2009). Activity patterns extracted from GPS data will allow us to examine travel behavior, traffic congestion, community participation, and human impacts on the environment at a fine granularity.

Whether this potential can be realized, however, rests on accuracy of algorithms that detect meaningful patterns from GPS track data. Unfortunately, GPS track data is largely characterized as voluminous and noisy. To illustrate, tracking the movement of an individual during a week period with a five second recording interval creates 120,960 track points. Positional accuracy of GPS measurements varies although the accuracy has improved significantly over the last two decades. It is hard to obtain complete coverage of reliable GPS tracks over an extended period of time; it is mainly due to signal loss especially indoors. This renders manual post-processing of GPS data nearly infeasible with large population over an extended period of time. While needs for the automated program exist, developing the algorithm robust to uncertainty remains to be a challenging task.

The purpose of this paper is to determine whether DBSCAN (Ester et al., 1996), a density-based spatial clustering algorithm, can be adapted to improve the performance of GPS data post-processing. The performance is defined as the ability to detect trips accurately from GPS data. A problem of detecting trips comes down to detecting stops that form an origin and destination of a trip. Stops are geographic events that individuals visit to participate in a certain activity, and provides a clue to individuals’ community participation. DBSCAN can be applied to detecting stops, as noisy spatiotemporal clusters. However, little attempt has been made to modify DBSCAN to detect significant stops, spatiotemporal clusters. The proposed method builds on Hwang et al. (2013), and we offer new findings with revised parameters.

2. Related work

GPS post-processing methods aim to extract activity patterns or a personal itinerary from GPS track data. A personal itinerary is more useful (i.e., parsimonious but more informative) than raw
GPS data. For the sake of operationalizing a problem, the itinerary will consist of two geographic events in different geometry types: a *stop* as a place that an individual visits in order to participate in one or more activities for some duration, and a *trip* as a route from one stop (origin) to another stop (destination) for some duration. A stop is Point and a trip is LineString in terms of Open Geospatial Consortium (OGC) geography data types.

Post-processing GPS data consists of three steps: data cleaning, trip detection, and trip characterization (Schuessler & Axhausen, 2009). Data cleaning treats track logs that lack consistency and accuracy that might negatively affect performance of data post-processing algorithms. Trip detection algorithms extract a route taken between a pair of origin and destination. Trip characterization algorithms assign attributes (e.g., trip length, mode of trip, purpose of trip) to a detected trip. This paper focuses on trip detection.

A common approach to trip detection is to specify rules based on relevant features using a series of if-then statement or decision-tree (Stopher et al., 2008; Schuessler & Axhausen, 2009; Srinivasans et al., 2009). For example, track points are classified into a driving (or walking) trip if speed meets a certain range of values. Similarly, a set of track points can be flagged as a stop if there is little change in position between consecutive track points. In a sense, trip detection is a classification problem that determines whether track points are trips or stops, leading to inference of activity patterns. Machine learning and filtering techniques have been applied to trip detection and characterization, including inferring mode of transportation (Zheng & Zho, 2011; Zheng et al., 2008). Although rule-based approaches described above are relatively easy to implement, it has been shown to be challenging to specify rules that are robust to uncertainty present in GPS track data.

Stops can be alternatively detected using density-based spatial clustering algorithms. DBSCAN, one of density-based spatial clustering algorithms, detects a group of dense spatial clusters by aggregating spatial clusters that are density-reachable (Ester, et al., 1996). DBSCAN begins with scanning the number of data points within a pre-specified bandwidth $\epsilon$ from any unvisited arbitrary data points (*seeds*). If the number of data points exceeds a pre-specified minimum number of points (*MinPts*), those data points within $\epsilon$ are set to a spatial cluster $C$. For each $C$, the algorithm checks to see if each point within $C$ forms another spatial cluster (i.e.,
density reachable). All density reachable data points are marked as clusters. Data points that are not density reachable are marked as noises.

DBSCAN can be used to delineate a stop as a spatial cluster, and noises that are not identified as a cluster will be set to a trip. While it is extremely difficult to distinguish between stops and trips particularly during a warm start (when GPS receivers regain signal reception after signal loss), it is advantageous that DBSCAN can overlook some noises in detecting stops. Significant places were detected using DBSCAN (Schoier & Borruso, 2011). This study, however, detects a significant place (a spatial cluster), not a significant stop (a spatiotemporal cluster). A spatial cluster does not differentiate when a place is visited whereas a spatiotemporal cluster does. To detect significant stops, it is necessary to modify DBSCAN such that a spatial cluster can be disaggregated by time of visits.

A few lessons can be learned from previous studies. It appears that rule-based approach alone cannot classify track points accurately. GPS track data contain some noise, and thus enforcing rigid thresholds without regard to uncertainty and contexts will not yield accurate results. Eclectic approaches that combine event detection algorithms robust to noise with mechanisms that take into account data quality and temporal contexts can produce reliable results. Below we describe the proposed method for automatically detecting trips from GPS data.

3. The proposed method

Once duplicate and anomalous records are treated, the proposed trip detection algorithm (a) identify candidate stops (spatial clusters) as input of DBSCAN using kernel density estimation (KDE), (b) modify DBSCAN to detect spatiotemporal clusters, and (c) smooth out any remaining misclassified track points using temporal filtering. An overview of the proposed method is shown in Figure 1.

Frequently visited places can be identified using KDE because those places have unusually high density values. Narrowing down candidate stop locations instead of random search reduces running time. Spatial clusters identified from DBSCAN are checked to see if those clusters are consecutive for at least 4 minutes. If this condition is met, spatial clusters are disaggregated into stops with different time of visits. For DBSCAN, \( \epsilon \) (bandwidth) is set to 100
meters, and $MinPts$ is set to 5; $\varepsilon$ is chosen based on observed spatial accuracy of data and extent of spatial clusters, and $MinPts$ is decided in relation to a recording interval; we use 30 second recording interval. With DBSCAN, track points that are scattered around stops but do not form part of a stop (e.g., beginning of a new trip) are correctly classified into trips. Finally, the program calculates a majority value from five consecutive track points, and assigns a majority value to those track points to smooth out misclassified values.

![Data Cleaning Diagram](image)

Figure 1. Process of clustering-based trip detection algorithmS

3.1 Data cleaning

Quality of raw GPS data was inspected and treated if necessary to minimize effects of uncertainty on trip detection algorithms. First, geocoded track points are overlaid on data of higher accuracy and independent source (ArcGIS 10.1 Map Service World Imagery). It was shown that most of GPS data were well aligned with reference data, but there were positional outliers (i.e., track points deviated from routine transportation routes). A level of uncertainty present in raw GPS data varies as data were collected in different time and weather conditions.
The program checks for any anomalous track points. Anomalous track points are detected by abrupt change in speed. More specifically, change in speed between two consecutive track points was calculated as change in location divided by change in time. A track point is deleted if the value (change in speed) is greater than 130 kilometer per hour. Duplicate records were checked because software proprietary to the GPS logger used for this study limits the number of track logs (points) that can be exported, which invites human errors during manual data export.

3.2 Event detection

3.2.1 Identify seeds for DBSCAN using Kernel Density Estimation

Instead of choosing seeds (unvisited arbitrary data points) randomly for DBSCAN, the program narrows down the location of candidate stops using Kernel Density Estimation (KDE). KDE scans the number and proximity of points to regularly placed location within a pre-specified bandwidth, and transforms that information to smooth kernel density function. Frequently visited places can be identified using KDE because those places have unusually high density values. For KDE, pixel size is set to 30 meters, and bandwidth is set to 90 meters. Extreme KDE values—used to identify locations of candidate stops—are defined as values greater than or equal to 900 per square kilometers.

Representative location of extreme KDE values is fed into DBSCAN as seeds; this serves to reduce processing time for DBSCAN, and minimize chance of missing potential stops due to random and incomplete search of the DBSCAN algorithm. To prevent the algorithm from being trapped in home (where vast majority of time is spent on), track points presumed to be at home location are excluded for analysis.

3.2.2 Detect stops using temporal DBSCAN

Spatial clusters identified from DBSCAN are checked to see if those clusters are consecutive for some duration of time. If the abovementioned requirement is met, spatial clusters are disaggregated into stops (as spatiotemporal clusters). So a place visited more than once are identified as multiple stops. Track points that are not stops are classified into trips.
For DBSCAN, $\varepsilon$ is set to 100 meters, and $\text{MinPts}$ is set to 5. 100 meters are chosen based on observed spatial accuracy of data and extent of spatial clusters. It should be noted that life space of study participants are largely suburban in U.S with distinct land use compared to other continents. For a spatial cluster to be qualified as a spatiotemporal cluster (that is stop), duration should be longer than 4 minutes, and track points making up a stop should be continuous. DBSCAN is selected to detect stops over threshold-based methods described earlier. With DBSCAN, track points that are scattered around stops but do not form part of a stop (for example, beginning of a new trip) are correctly classified into trips.

3.2.3 Smooth out misclassified values using temporal filtering

Stops and trips identified based on a density-based spatiotemporal clustering (as described above) can still remain misclassified. In particular, temporally noisy values can exist quite often. For example, a series of track points have an array of values [stop, trip, stop, stop, stop] although they should be classified into [stop, stop, stop, stop, stop]. To resolve this problem, the program calculates a majority value (the most common value) from consecutive track points with total duration 2 minutes and 30 seconds, and assigns a majority value to those track points. That way, an array of five track points [stop, trip, stop, stop, stop] can be classified into a stop as a whole. This effectively removes noise that remains after clustering-based classification.

3.2.4 Extract geographic events

Unique identifiers (IDs) are assigned to stops and trips based on temporal order and a rule that a stop is followed by a trip, and vice versa. In operational terms, a stop is a collection of track points that are spatially clustered and temporally continuous; a trip is a collection of track points that are not spatially clustered and temporally continuous.

Track points that are classified into stops and trips with unique IDs are aggregated (or dissolved in GIS terms) into two geographic events called stop centers and trip routes. Stop centers are mean centers of track points with a unique stop ID; trip routes are polylines that link an origin stop to a destination stop (Figure 2). Figure 2 shows stop centers and trip routes extracted from GPS data using the proposed algorithm. It can be seen that this subject made eleven significant stops during the week period. Stop centers and trip routes are labeled by IDs that follows temporal order. Staying at home is not included in the map.
4. Evaluation

The trip detection algorithm described above was applied to data collected for the study that monitors mobility of subjects using GPS (Evans et al., 2012). Total 57-week worth of data from 17 subjects was collected. Subjects are provided an informed consent to participate in the study, involving carrying a GPS logger (GlobalSat DG-100 Data Logger) during waking hours of each week for this study. Recording interval was set to 30 seconds, and data were collected in a passive and continuous mode.

All GPS track data were collected in one week time period. We collected data amounting to total seven week periods extending from May 2011 to March 2014. Subjects are divided into control subjects and rehabilitation subjects during week 1, week 5, week 9, month 6, and year 1.
after treatment. For each week data amounted to about 5,000-27,000 track points depending on
the extent of gaps in the data. Data was processed using the trip detection program written in
Python 2.7 with ArcGIS 10.1.

To evaluate how accurately the proposed method classifies track points into stops and
trips, we check classification errors of 30% sample against results of manual classification. For
manual classification, we manually check all track points to see whether they constitute stops or
trips by looking through all track points superimposed over ArcGIS 10.1 Map Service World
Imagery.

Table 1 shows an error matrix compiled from twenty data sets. The percent correctly
classified (PCC) is 93.8%, and Kappa index, a measure of classification accuracy, is 0.87. That is,
data are correctly classified 87% of the time with the proposed method taking into account the
agreement occurring by chance.

<table>
<thead>
<tr>
<th></th>
<th>Trip</th>
<th>Stop</th>
<th>Row Total</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Trip</td>
<td>328</td>
<td>25</td>
<td>353</td>
<td>PCC:</td>
<td>93.80%</td>
</tr>
<tr>
<td>Stop</td>
<td>10</td>
<td>200</td>
<td>210</td>
<td>Kappa:</td>
<td>0.87</td>
</tr>
<tr>
<td>Column Total</td>
<td>338</td>
<td>225</td>
<td>563</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Conclusion

The proposed method utilizes temporal DBSCAN for detecting stops and temporal filtering for
smoothing misclassified track points. The method classifies GPS track points into stops and trips
with high level of accuracy (87%). Results show that a density-based spatiotemporal clustering
combined with temporal filtering is effective in detecting stops and extracting the personal
itinerary from GPS track data.

This study demonstrates that DBSCAN can be adapted to differentiate between two types
of geographic events—significant stops and trips between stops—in GPS track data. Further,
density-based spatiotemporal clustering (or temporal DBSCAN) combined with temporal
filtering can better handle noise present in data, and thus improve inference accuracy. Being
robust to noise is advantage of the proposed method over threshold-based methods (such as attribute query).

Several limitations of this research should be acknowledged. The proposed method accurately classifies track points within the constraints of the total time for GPS data collection. In other words, accuracy of the method is not assessed with respect to gaps in data. Gaps in GPS data can be filled using other complementary means of measuring mobility. Research shows that combining accelerometry-based data with GPS data can improve reliability of mobility measurement (Oliver et al., 2010).

The performance of the method can be evaluated using more rigorous methods. This research examines combined effects of clustering and filtering on inference accuracy. That is, it does not examine independent effects of clustering and filtering, respectively. The performance of the proposed method can be compared against more rigorously formulated threshold-based methods in different variation. The proposed method can be also evaluated in reference to self-reported information completed in real-time rather than in reference to manual classification results.

Meaningful stops extracted from the program can be used to analyze activity patterns and community participations of individuals. Similarly, meaningful trips (with mode of transportation) will be useful in understanding transportation mode choice, and estimating transportation-related carbon footprint of individuals.

References


Big Data, Small Apps: Premises and products of the civic hackathon

Sara Jensen Carr
PhD candidate
Department of Landscape Architecture and Environmental Planning
University of California, Berkeley
202 Wurster Hall #2000
Berkeley, CA 94720-2000 United States
secjensen@berkeley.edu
(504) 220-2117

Allison Lassiter
PhD candidate
Department of Landscape Architecture and Environmental Planning
University of California, Berkeley
202 Wurster Hall #2000
Berkeley, CA 94720-2000 United States
abl@berkeley.edu
(510) 859 3276
Abstract
Connections and feedback among urban residents and the responsive city are critical to Urban Informatics. One of the main modes of interaction between the public and Big Data streams is the ever-expanding suite of urban-focused smartphone applications. Governments are joining the app trend by hosting civic hackathons focused on app development. For all the attention and effort spent on app production and hackathons, however, a closer examination reveals a glaring irony of the Big Data age: to date, the results have been remarkably small in both scope and users. In this paper, we critically analyze the structure of The White House Hackathon, New York City BigApps, and the National Day of Civic Hacking, which are three of the most recent, high-publicity hackathons in the United States. We develop a taxonomy of civic apps, analyze hackathon models and results against the taxonomy, and evaluate how the hackathon structure influences the apps produced. In particular, we examine problem definitions embedded in the different models and the issue of sustaining apps past the hackathon. We question the effectiveness of apps as the interface between urban data and urban residents, asking who is represented by and participates in the solutions offered by apps. We determine that the transparency, collaboration and innovation that hackathons aspire to are not yet fully realized, leading to the question: can civic Big Data lead to big impacts?

Keywords
app, hackathon, participation, representation, open governance
Introduction

In the age of Big Data, mobile technology is one of the most crucial sources of data exchange. Analysts are examining the preferences, behaviors and opinions of the public through status updates, tweets, photos, videos, GPS tracks and check-ins. In turn, urban residents are accessing the same data as they view restaurant reviews with Yelp, find a ride with Uber, locate their friends with Foursquare, and stream Instagram photos. Untethered devices such as smartphones and tablets are critical to real-time, on-the-go data uploading and access. On these mobile devices, the public is connecting to data through apps. Apps are becoming the primary interface between data and the public.

At present, apps are primarily created by private companies seeking to profit from granular knowledge of urban behaviors. Yet, the allure and potential of apps is increasingly recognized by non-profit and government organizations, with development encouraged from the federal government all the way down to local municipalities. In the private tech industry, the “hackathon,” a short and intense period of collaborative brainstorming, development, and coding, is a standard model of app development, and now the public sector is following suit. So-called “civic hackathons” are rapidly proliferating. New York City’s BigApps contest is in its fourth year, the White House recently hosted its second annual hackathon, and the 2014 National Day of Civic Hacking, in its second year, claims a 30% increase in events from its first hackathon (NYCEDC 2014; Heyman 2014; Llewellyn 2014).

Many view civic hackathons and app development as an exciting indicator of a new era of collaborative, open governance and bottom-up engagement. In the words of the promoters behind the National Day of Civic Hacking,

*Civic hackers are community members (engineers, software developers, designers, entrepreneurs, activists, concerned citizens) who collaborate with others, including government, to invent ways to improve quality of life in their communities...Participants will use technology, publicly available data, and entrepreneurial thinking to tackle some of our most pressing social challenges such as coordination of homeless shelters or access to fresh, local, affordable food.* (Hack for Change 2014a)

In private industry, the utility of the hackathon are usually clear: employees work to innovate new products that will keep the company on the cutting edge of the market, often with potential
shared profits (Krueger 2012). The goals of civic hackathons are less so. Ostensibly, they consist of citizen developers and representatives donating their time to create apps that address community wants and needs. The structure of the events, however, heavily influences the kind of apps produced, their intended users, and their long-term sustainability. In this study, we examine three models of civic hacking, develop a taxonomy of civic apps based on their structure, and offer cautions and suggestions for future civic hackathons.

**Premises of the Hackathon**

Hackathons first gained popularity through the 2000s, as technology companies informally hosted hacking marathons. Hackathons were intended to promote exploratory coding, new idea generation and prototyping in a low-risk environment. As a Facebook software engineer Pedram Keyani wrote, "hackathons are our time to take any idea—big or small, sane or crazy—and build it into something real for people to react to" (Keyani 2012). Facebook held its first official, in-house hackathon in 2007. Since then, "[e]very couple of months, a few hundred of our engineers unleash their talents in epic, all-night coding sessions," working alongside people in different departments with different skill sets. Facebook's hackathons were greeted with the intensity and enthusiasm common among young coders, where "...everyone keeps working until around 6:00 am or when they pass out—whichever comes first" (Keyani 2012). Several of Facebook's key features, such as the Like button, were born out of hackathons (Krueger 2012).

Since the first all-night pizza- and caffeine-fueled marathons, hackathons have expanded and become more inclusive. Facebook now hosts daytime events, allowing a wider range of people to participate, including potential recruits (Krueger 2012). Other companies opened hackathons to the public. In 2013, Salesforce hosted a hackathon that challenged anyone to creatively use their data and APIs in a mobile app, rewarding the winning entry with $1 million, "the biggest single prize in hackathon history" (Salesforce 2013). Prize money is sometimes used as an incentive for participation, although not always. In some cases, people voluntarily participate out of personal enthusiasm, curiosity, or attachment to the hacking community, which is certainly a boon to companies hosting the events.

Not surprisingly, the low cost, low risk, and often innovative hackathon model is now spilling out beyond the tech industry. For governments, hackathons seem to offer answers to two major
issues. Firstly, governments often have a wealth of data, yet lack the capacity to process and innovate with it. Hackathons not only bring in virtually free labor in the spirit of "volunteerism and civic duty," (Hack for Change 2014b) but often facilitate mobile app development. Supporting app development helps governments seem contemporary, innovative and efficient. Moreover by releasing the data used in app development, governments can make the claim to transparency in their operations.

Since participation and team collaboration are built into hackathons, they foster a spirit of open governance. Citizens define their own problems and solutions, as well as a voice for their own communities, often while working with government-supplied Open Data. Material for the National Day for Civic Hacking lauds that hackathons are "... representative of a movement that is underway to leverage the power of technology and engaged neighbors to minimize barriers between government and citizens. National Day of Civic Hacking is truly about citizens stepping up to their role in government from the local to the state to the federal level...NDoCH represents the movement toward a truly collaborative government/citizen relationship of the future" (Hack for Change 2014b). Whether the hackathons have lived up to this promise has yet to be seen.

Hackathons professes a more fluid democracy, a seeming rebuke to the top-down problem solving of the previous age of urban informatics (Greenfield 2013, Townsend 2013). As an article on the Open NASA blog states, "hackathons frequently show us insights and applications that we never would have imagined coming from our own work. These technology development events don’t give us all the answers – but they engage the public in exploration of our data and our challenges in creative and compelling ways, sparking a flame that just might become something big and powerful. A hackathon isn’t a product, it’s an approach..." (Llewellyn 2012); an approach intended to engage the public and harness the creativity of the crowd. This focus on process, however, is often neglected by the administrators of these contests, who tout its potential to problem solve and tackle challenges (NYCEDC 2014; Hack for Change 2014a).

As Open and Big Data proliferate, there is seemingly little reason not to hold a hackathon. If there is no prize money on the line, the stakes can be very low. However, to use a tech industry term, are hackathons “disrupting” models of governance by widening participation? Is app generation leading to more efficient problem solving?
Civic Hackathon Models

In the United States, three of the most visible, recent hackathons are the 2013 White House Hackathon, the 2013 New York City BigApps contest, and the 2014 National Day of Civic Hacking. These three hackathons embody the different approaches and goals of civic hackathons. The scales, organizational principles, and predominant types of apps generated vary, but together they reveal issues common across hackathon models.

API Demonstration at the White House Hackathon

Since the Obama Administration took office, they have supported open data and open governance. Their agenda was formalized through 2009's Open Government Directive. The directive required all 143 United States federal agencies to upload all non-confidential data sets to a newly created sharing platform, data.gov, by November 9, 2013 (White House 2009). Disseminating newly disclosed data through a hackathon is attractive, because of the transparency and participation associated with the model. Hackathon proponents assert that creating civic apps is critical to maximally leveraging Open Data.

The 2013 White House Hackathon focused on an Application Program Interface (API), a code library that gives immediate access to data as it is updated. The API, We the People, provides data on federal petitions, such as information concerning when the petition was created and how many people have signed it. From this API, civic hackers developed apps that made the data accessible in other formats, show the number of signatures in real time, and map spatial patterns of support (The White House We the People 2013). Given the limited data at hand, the fact that over 30 apps developed could be considered a success. However, of the 66,146 available datasets on data.gov, which range from environmental to budget information, the choice to make petition information the focus of the first hackathon is puzzling. Possibly, petition data were selected as the locus because petitions indicate an open, collaborative government, but creating better access to petition data does little to change the minimal impact of petitions in federal governance. The second White House hackathon held in 2014 (results still pending at the writing of this paper) continued to focus on petitions with Petitions Write API, which is intended to expand the platforms and sites through which people can submit petitions (Heyman 2014).

The 2013 White House Hackathon produced many prototypes, but few enduring projects. Since the competition, none of the apps have been institutionalized by the government, nor does it
appear the apps have been updated or distributed. Video demonstrations of the apps are available, but few of the apps are directly accessible or downloadable. One downloadable program, a code library that extends analytic possibilities by porting petition data into the statistical program R, does not work with the current version of R.

While the initial tenets of Obama’s Open Government Directive were “transparency, participation, and collaboration,” (White House Memo 2009), the collaboration component has largely fallen by the wayside (Peled 2011). Spokespeople for the US government’s Open Data Initiative now choose to highlight the potential of transparency and downplay the failure of federal agencies to use their data. The White House has made a very public push to tout transparency as a virtue in of its own sake. While the hackathon may widen the range of participants in governance, it falls short of deeply collaborative, open governance.

One of the tangible successes of the hackathon is that all the code produced was made available through the public code repository, GitHub. Everything that was finalized in the hackathon is now a public resource, so it could potentially be accessed and built upon in the future. The optimistic view is that “[w]ith each hackathon, some of the detritus — bits of code, training videos, documentation, the right people trading email addresses — becomes scaffolding for the attendees of later ones” (Judd 2011). It may be left to the developer community, however, and not the White House, to expand the results.

**App Competition at New York City’s Big Apps**

New York City’s Big Apps contest is one of the longest running civic hackathons. It is also considered one of the most successful, in terms of apps sustained past the contest period. The competition origins lie in a large push from former Mayor Michael Bloomberg, who also pioneered some of the first Open Data legislation in the country and was the first to appoint a Chief Information and Innovation Officer, Rahul N. Merchant, in 2012 (New York City Office of the Mayor 2013). Both Bloomberg and Merchant brought considerable experience in the business, financial and technology sectors. They were able to garner major sponsors such as Facebook, eBay, Microsoft, and Google for the contest.

Some believe this kind of private-public partnership is key to garnering talent and funding apps that will sustain beyond the competition. In 2013, a panel of judges awarded $55,000 to the
BigApps winner, and amounts ranging from $5,000 to $25,000 to runners up in several categories. In all three years of the competition, BigApps has also always awarded an “Investor’s Choice” prize, underscoring its focus on apps that offer a financial return.

 Yet, BigApps does not explicitly require financial returns. Instead, it asks that participants explore ways to use technology to make New York City a better place to “live, work, learn, or play” (NYCEDC 2014). It also offers participants a chance to tackle known challenges supplied by 30 private and public entities. These range from improving health access, finding charging stations for electric cars, and helping network parents of students in the New York City school system. Past participants have occasionally taken on these challenges, but most winners of the contest come up with their own ideas, usually with the intent of monetization. The first place winner of the 2013 contest, HealthyOut, uses a Yelp API to help connect people with healthier food delivery options. HealthyOut has since raised $1.2 million in venture capital, beyond the $55,000 in prize money earned from BigApps (Lawler 2013).

 The apps that have endured since the 2013 competition were able to secure substantial funding and develop a revenue stream. Hopscotch, an educational coding iPad app for children, also raised $1.2 million in seed funding (Lomas 2014) and is a top seller in the iPad store. The majority of BigApps winners with more community minded goals are no longer functional, since they were unable to create a sustainable financial model. HelpingHands, one of the prize winning apps that helped NYC residents enroll in social services, was not available in app stores at the time of this writing and the domain name was for sale.

 It could be argued that BigApps encourages entrepreneurship and feeds money back into the city. The leaders of BigApps claim successful ideas will be rewarded with the resources to sustain them (Brustein 2012). However, its additional claims “that [it] empowers the sharpest minds in tech, design, and business to solve NYC's toughest challenges” (NYCEDC) rings hollow. The slogan itself recognizes that only a select group is empowered. Perhaps as a course correction to this issue and response to several public criticisms of the non-civic goals of the winners (Brustein 2012), the 2014 contest will offer contestants a chance to work directly with public agencies to solve their specific, pre-defined problems.
Locally-driven Innovation at the National Day of Civic Hacking

Unlike the White House Hackathon and BigApps, the National Day of Civic Hacking is not a government-driven initiative. The hackathon is organized by consultant Second Muse with non-profits Code for America, Innovation Endeavors and Random Hacks of Kindness. It is sponsored by Intel and the Knight Foundation, with support from the White House of Science and Technology Policy, several federal and state agencies, and other private companies. Though geographically the largest and most widespread movement, NDoCH aims to address hyperlocal issues by promoting a coordinated set of nation-wide hackathons hosted by localities on their own terms (Hack for Change 2014c; SecondMuse 2013).

NDoCH has few rules and the hackathon is interpreted broadly by participating cities and states. The full set of projects from NDoCH is messy, but successfully conveys the varying interests in participating areas. Most groups created apps or websites, such as Hack 4 Colorado's FloodForecast, which notifies users if their home address is in danger of flooding. Other groups worked on alternative technical projects. Maine Day of Civic Hacking, for example, focused on repairing a stop motion animation film in a local museum.

Though not especially consistent with its local focus, NDoCH also promotes national issues. In 2014, formal Challenges were advertised by federal agencies like the Consumer Financial Protection Bureau and the Federal Highway Administration. The Peace Corps, for example, requested "a fun, engaging and easy-to-use interface with the numerous and diverse Peace Corps volunteer opportunities that helps the user find the right opportunity for them" (Hack for Change 2014d), which was subsequently prototyped at the San Francisco Day of Civic Hacking (Hack for Change 2014e). NASA's Challenge to increase awareness of coastal inundation spurred several related projects.

Like the White House Hackathon, the route between hack and implementation is unclear. But, NDoCH's guiding principles are oriented toward the process of the hackathon, rather than the results. Stated goals include "Demonstrate a commitment to the principles of transparency, participation, and collaboration" and "Promote Science, Technology, Engineering and Mathematics (STEM) education by encouraging students to utilize open technology for solutions to real challenges" (SecondMuse 2013). Participants may not be clear, however, that they are contributing to a broader process and most apps will not survive past the hackathon. This tension
is exacerbated by reports from the NDoCH that tout the number of apps produced, not just process-related goals. Ensuring authentic, collaborative processes over app development remains a challenge.

A Taxonomy of Civic Apps

From the outcomes of each hackathon, we delineated five categories that described the apps produced across all events. We examined the most recent results for each hackathon: White House 2013, New York BigApps 2013 and the National Day of Civic Hacking 2014. We evaluated their descriptions, demonstrations, and found them in the Apple and Android stores, as applicable. For White House and BigApps, we evaluated information on the winning entries. For NDoCH, there were no selected winners and information on associated efforts was available, so we examined all entries.

Spatial Customization and Personal Services

Many of our most used commercial apps such as Google Maps, Yelp, or Foursquare, specialize in the spatial customization of individual daily routines. These apps, powered by ever expanding GPS technology, have advanced the “spatially enabled society,” where citizens are better able to communicate with the world around them (Roche et al. 2012, 222). In spatially enabled society, "the question ceases to be simply 'Where am I?' and becomes: 'What is around me?' (as in services, people, and traffic), 'What can I expect?' and 'How do I get there?'” These apps center around easing mobility and in many cases consumption in the city – finding parking, giving real-time transit alerts, or customizing personal routes given a set of favorable inputs.

It is no surprise that many civic apps have also seized on expanding the suite of spatial customization apps, as mobility and movement in the city is a continuing challenge, while self-locating with GPS remains a relatively new possibility. The majority of the spatial customization apps provide real time transit alerts. In addition, services like Healthy Out tells a user which restaurants near them are best suited to their personal diet, while Poncho tells what the weather will be at every location in your daily routine. These spatial customization apps are focused on the desires of the individual, harnessing open data to ease everyday life.

Some of these apps also crowdsourced data from users, aggregate the data, and then provide users with continually updated, socially-derived urban data. Some of these apps are oriented toward
typically underserved populations. Ability Anyware Assistive Technology Survey and Enabled City, built during the NDoCH, identifies and find accessible routes and buildings for people with disabilities.

We also include personal services in this category. Even though this information is sometimes aspatial, these services do help make urban life more efficient to the individual. Two apps built at NYC BigApps, ChildcareDesk and HiredinNY, intended to connect users to child care centers and jobs, respectively.

*Spatial Awareness and Data Communication*

The spatially enabled society is also at the heart of many apps that are not expressly built for individual easing, but rather to simply visualize otherwise invisible information to increase awareness among app users. Apps produced at the Virginia Beach hackathon under NDoCH mapped the effects of potential sea level rise. Others mapped child hunger statistics (Maine Child Hunger Cartogram Viewer) and SNAP benefits (SNAPshot) with the professed intent of spurring empathy.

In addition, this category of apps uses spatial information to encourage individuals to act in their own community, by bringing visibility to difficult-to-perceive issues. Freewheeling NC, an app built at a North Carolina hackathon during the National Day of Civic Hacking, crowdsources bike routes with the intent of influencing urban planning. Many of these apps use public data on underutilized vacant land in order to help community members to use them as gathering places, urban agriculture, or new development (Minimum Adaptable Viable Urban Space (MAVUS); [freespace] ATX; Abandoned STL).

Some apps similarly raise awareness through aspatial data communication. Often aimed at government transparency, such as several apps developed at the White House Hackathon with the We the People API, or campaign finance information and city council agendas. At times, data communication and spatial awareness come together, such as Flood Forecast's flood notifications.

*Community Building*

Many apps built during the National Day of Civic Hacking focused on community building through peer-to-peer communication. These apps help people in niche groups find each other,
such as connecting pet owners and teens. Community building apps also help pair volunteers with nonprofit organizations such as Habitat for Humanity, help people find places to donate leftover food, and allow citizens direct access government officials.

Educational
The smallest group of apps are educational, which are often aimed at youth and incorporate a gaming component. The aforementioned Hopscotch at NYC BigApps teaches children to code, and two apps built under NDoCH aimed to teach users about watersheds and urban geography by using a Minecraft-like interface.

Data Gateways
Lastly, some apps are simply focused on making data accessible in a different machine-readable format or providing analytic environments for the data, but not for specific purposes. These are often interfaces geared towards developers to create even more apps. Almost a third of the apps from the White House Hackathon fall into this category; other data gateways were created for Peace Corps project data, hospital discharge costs, and even presidential inaugural addresses for textual analysis.

Models and Results
The structure, funding, and data released in each type of hackathon not only influenced the scale and number of apps produced, but the predominant type of apps. Because of the large variations in end product numbers – 17 for the White House Hackthon, 7 in the BigApps contest, and 71 during the National Day of Civic Hacking – the percentage of the total apps is given in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Spatial Customization and Personal Services</th>
<th>Spatial Awareness and Data Communication</th>
<th>Data Gateways</th>
<th>Community Building</th>
<th>Educational</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>White House Hackathon 2013</td>
<td>0% (0)</td>
<td>71% (12)</td>
<td>29% (5)</td>
<td>0% (0)</td>
<td>0% (0)</td>
<td>0% (0)</td>
</tr>
</tbody>
</table>
Table 1. Results from the White House Hackathon 2013, New York City's BigApps Contest 2013, and the National Day of Civic Hacking 2014. The total number of apps produced is given in parentheses.

<table>
<thead>
<tr>
<th>Event</th>
<th>Apps</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYC BigApps 2013 (7)</td>
<td>86% (6) 0% (0) 0% (0) 0% (0) 14% (1) 0% (0)</td>
</tr>
<tr>
<td>National Day of Civic Hacking 2014 (71)</td>
<td>13% (9) 34% (24) 10% (7) 28% (20) 8% (6) 7% (5)</td>
</tr>
</tbody>
</table>

At the White House Hackathon, no winners were declared, but results and code were posted for only 17 out of 30 completed projects executed at the hackathon. Of the 17 apps, 12 communicated Spatial Awareness and Data Communication, while 5 focused on Data Access. This is no doubt related to the focus on only one dataset. It is also notable that the White House hackathon goals were not to solve any particular urban challenge, but rather simply to see what technological expertise could do with a set of open data.

The total number of entries in New York City's BigApps is not available, but all 7 winning entries earned prize money. It was the only hackathon of the three to offer prize money and the only to boast significant private business partners and potential for investors. Among the apps, 6 out of 7 are focused on personal mobility and personal services, with the remaining app was educational. While the contest claims that it is bringing together experts and developers to “solve New York’s toughest challenges,” its results thus far indicate that it is more focused on apps that ease individual consumption and mobility; instead of public data sets, the winning entry used the commercial API from Yelp.

While the White House Hackathon and BigApps each produced narrow results, the National Day of Civic Hacking generated many apps, crossing all five categories. Of the 71 products from the competition, some of the results are as technically sophisticated as those developed in BigApps and the White House Hackathon. Others entries are not apps at all – they are requests for apps or brainstorming sessions regarding the potential for apps. These "Other" results make up 7% of all the entries into NDoCH. Notably, however, 28% of NDoCH projects focused on Community Building, which was absent from the other two hackathons. Of Spatial Customization and Personal Services apps, several focused on finding accessible facilities for disabled citizens or improving public amenities such as bike lanes, trails, and parks. As the NDoCH organizers
hoped, the apps associated with NDoCH reveal community-driven, locally-specific issues and civic innovation. However, they also reveal the technological limits of many localities. It is not surprising that the greatest number and most sophisticated apps come from technology hubs such as Palo Alto and Austin, where some of the more distant outposts did not have the expertise to even produce an app at the end of the event.

**Future Hackathons**

The landscape of civic data and civic apps is rapidly changing, corresponding with the rise of Open Data and Big Data, expanding mobile technology, and trends toward technocracy (Mattern 2013). Promoting civic hackathons is not only a low-risk, adaptable method of embracing contemporary problem solving amidst change, but previous ones have generated enough success to keep attempting them. Hackathons have fostered new types of civic participation, created enthusiasm among some community members, developed some new apps, and may broadly encourage more technological innovation in government. Yet, while there is evidence of successes, civic hackathons face unique challenges that must be addressed in order to deepen their impact.

The examples proffered here show that there are three common and interrelated issues with the hackathon model. Firstly, defining problems that are meaningful for the community. Secondly, the challenge of aligning the goals of market-ready apps with civic services. Thirdly, and most importantly, that the civic hackathon has the responsibility of addressing the needs of its full constituency, not simply the smartphone-owning, tech-literate public.

**Defining Problems**

Many experts have identified that lack of structure in hackathons, meant to encourage out of the box approaches, can lead to unfocused results. NASA's open data portal, for example, states that the key to implementing a successful hackathon "is to invest the effort to identify the right problem statements, provide the supporting data, and get a good mix of people in the room" (Llewellyn 2012). As data scientist Jake Porway warns, "They are not easy to get right... You need to have a clear problem definition, include people who understand the data not just data analysis, and be deeply sensitive with the data you’re analyzing" (Porway 2013).
When any local government is faced with a challenge, deeply understanding the data being analyzed, the physical and social context, and possible opportunities is crucial. Developing the web of knowledge necessary to solve most issues takes time. While the public can offer needed fresh perspectives and local insight, it can be difficult for outsiders to come in and hit the ground running, which is necessary in a short-lived, intense hackathon. Many of the apps that come out of these contests help users navigate the city, rate local places, or plan itineraries – all services that are already well-covered and arguably better developed by large tech companies (Brustein 2012).

Yet, “problem solving” is often touted as the key tenet of many hackathons. BigApps seeks to "solve specific New York City challenges, known as BigIssues" (New York City Office of the Mayor 2012). BigApps identified four BigIssues for the 2013 competition: Jobs and Economic Mobility, Healthy Living, Lifelong Learning, and Cleanweb: Energy, Environment, and Resilience. Of course, it is nearly impossible to solve any sort of complex issue, like a BigIssue, in the context of a hackathon. Creating nuanced solutions requires both quantitative expertise and experience. If hackathons are going to meaningfully address difficult city problems, it likely is necessary to create hybrid teams of public participants and government employees and commit to working through hackathon proposals beyond the short-lived timeframe of the hackathon.

Market-Ready Solutions
Because the hackathon is temporary by definition, perhaps it is not surprising that the results are commonly temporary as well. Hackathon products are often still in the brainstorming stage and are rarely taken to completion. The bulk of the apps that survive past a hackathon are market-ready and able to attract venture capital during or shortly after the hackathon.

At NDoCH, the organizers recognize that governments will not bear the responsibility for apps after their creation:

“Each new technology has a unique path to implementation. The key to the development of technologies that make their way out of the hackathon environment and into your community are public and private partnerships. One path to sustainability is that a group of volunteers develops a new app to connect low-income residents to the nearest free tax preparation site over the course of National Day of Civic Hacking. Following the event, the volunteers reach out to economic justice groups in their community so they can promote their services
using the app, seek a sponsor to offset the cost of the text message usage, and work with government officials to promote the app as well as the availability of free tax prep in your city.” (Hack for Change 2014b)

Not only is the onus of innovation and development shifted to a narrow swath of the data-literate public, but the growth and sustainability is as well.

Sustainability is identified as the primary issue by many technologists. Code for America's Dan Melton writes, "...some of the biggest examples of disconnect and potential opportunity come out of app contests or hackathons. Policy makers/political leaders champion city or social contests, to which, developers respond with dozens or even hundreds of submissions. So far so good. When the app contest is over, often too is the partnership" (Melton 2011). O'Reilley Media editor Andy Oram adds, "...how could one expect a developer to put in the time to maintain an app, much less turn it into a robust, broadly useful tool for the general public?...The payoff for something in the public sphere just isn’t there" (Oram 2011). Organizations like CivicApps (http://civicapps.org/), help to overcome sustainability issues by promoting apps for wider distribution, but nonetheless few are self-sustaining. Even the more civic-minded winning apps at NYC BigApps 2013 such as Helping Hands and HiredinNY are nowhere to be seen one year later, despite funding awards.

There is some implication that failure may be, in part, because of app quality. Refining existing ideas could help improve sustainability. Joshua Brustein of the New York Times (2012) says, “Inevitably, most of these projects will fail. Start-ups usually do. And considering the modest price of the program — BigApps costs the city about $100,000 a year, according to the city’s Economic Development Corporation — the bar for success should be set low. But it seems that a better investment might be to spend more time working with a few developers on specific ideas, rather than continually soliciting new ones” (Brustein 2012). Tackling this issue will require hackathon organizers to turn an eye to building communities before building apps. Clay Johnson of the Sunlight Foundation, a nonprofit dedicated to government transparency, notes that they see their hackathons as only the beginning of their engagement with both developers and volunteers (Johnson 2010). There are some cases of longer term partnerships, such as the Federal Registry's partnership with the winners of the 2010 Apps for America civic hackathon to create a
redesigned data distribution portal (Oram 2011), but these examples of government commitment are rare.

**Participation and Representation**

Revenue and consumption are the bases of most of the successful, sustaining hackathon propositions. This makes it challenging to address problems that are not profitable. Issues of the "unexotic underclass," such as veterans and welfare recipients, often go unaddressed (Nnaemeka 2013). This is evident despite Open Data's promise of egalitarianism and the participatory goals of hackathons.

One piece of this challenge is the demographic of participants that hackathons typically attract. The majority of hackathon developers are young, well-educated, and relatively affluent. Unsurprisingly, the majority of apps cater to this demographic, even in a civic context. As technologist Anthony Townsend (2013, 166) writes, “...should we be surprised when they solve their own problems first?...Not only do they not represent the full range of the city’s people; often these hackers lack a sense that it’s even their duty to help others...”. Representation bias is also noted by Porway (2013), when he similarly writes about a New York City hackathon focused on greening the city, "...as a young affluent hacker, my problem isn’t improving the city’s recycling programs, it’s finding kale on Saturdays."

Organizers should particularly look to increase the participation of women and low-income communities. The all-night structure and lack of code of conduct can be intimidating to women (Rinearson 2013). The National Center for Women and Information Technology notes the importance of specifically recruiting girls for events, not only as coders but as judges and mentors (NCWIT n.d.). One organization, Yes We Code, is responding by hosting their own hackathons that support ideas from low-income teens (Yes We Code 2013). Rather than hosting separate events, however, it is the responsibility of the city to ensure that such organizations and their constituents have a voice in the city-sponsored hackathon.

Representation bias also extends to the hardware of apps themselves. Only 56% of the U.S. population own smartphones and owners are primarily people under 35, well educated, and affluent (Smith 2013). Apps are, themselves, a limiting format. Though they continue to increase in popularity, many difficult to reach groups cannot be accessed with apps. In a world where
mobile data contributes to visibility and voice, those that are not able to partake become invisible.

Despite claims to openness, accessing and utilizing data still requires a certain amount of education and associated financial investment that is not available to large portions of the population (Gurstein 2011). This digital divide means that sometimes apps, with even the best intentions, can inadvertently exacerbate power structures. In San Francisco, an app called Handup used a government database on San Francisco’s homeless population to allow people to donate money for “good uses” to the homeless (Packer 2013). Unlike simple spatial awareness apps which can help people find food banks or donation centers, Handup, which essentially asks its users to pass value judgments on who and who does not merit charity, is an example of how apps can enforce a power structure. Only the privileged who own the technology make decisions about individual assistance.

Recent language in the NYC BigApps contest and National Day of Civic Hacking acknowledges that apps can be exclusive. In 2014, BigApps will accept competition entries that use a wider array of technology products (NYCEDC 2013). Expanding the technology to gather data and input may have surprising outcomes. When New York City's non-emergency reporting system, 311, added a website and then app to their telephone reporting system, they expected call volumes to go down. Instead, they saw an overall increase in reporting, showing that different mediums helped access different sectors of the population. Moving away from the app interface could encourage broader engagement.

**Conclusion**

While the hackathon has been a reasonably convenient solution thus far, the existing hackathon models make it difficult to address truly complex, non-monetizeable issues. As governments grapple with what to do with their newly opened datasets and how to handle much of their Big Data, they are left with some difficult challenges. How should governments ensure that civic innovations are institutionalized and sustained without being dependent on private backers? How do they ensure everyone is being fairly represented – that those on the far side of the digital divide are not left out of the wake of technological progress?
The first step to improving the civic hackathon is to subject it to the same scrutiny as any other urban practice. This includes clarifying the goals of hackathons and developing associated metrics. If apps have the possibility of creating efficiency gains, results should be internalized and governments should commit personnel resources to hackathons, support scaling, and dedicate money for ongoing operation. Alternatively, are hackathons intended to kickstart for-profit app businesses? If so, the role of public money in this process should be made clear and the surface claims to solving complex urban issues should be eliminated. Or, are hackathons a method of signaling open governance? If so, this method of participation should be examined against existing models of collaborative governance. While the analytical literature surrounding hackathons is scarce, it is necessary to develop best practices for running a hackathon and building on the results. In doing so, the hackathon may have the opportunity to become everything it wants to be: transparent, collaborative, and innovative. For now, however, the lofty goals remain unmet.

Sources


Greenfield, A. (2013). Against the smart city (the city is here for you to use) [Kindle eBook ed.] Amazon Digital Services, Inc.

Gurstein M.B. (2011). Open data: Empowering the empowered or effective data use for everyone? First Monday First Monday, 16(2).


‘Big Data’: Pedestrian Volume Using Google Street View Images

Li Yin¹ Qimin CHENG² Zhenfeng Shao³ Laiyun Wu⁴

¹Department of Urban and Regional Planning, University at Buffalo, The State University of New York, Buffalo, NY 14214, USA. Email: liyin@buffalo.edu
²Department of Electronics and Information Engineering, Huazhong University of Science & Technology, Wuhan 430074, China
³The State Key Laboratory of Information Engineering on Surveying Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China
⁴Department of Urban and Regional Planning, University at Buffalo, The State University of New York, Buffalo, NY 14214, USA. Email: laiyunwu@buffalo.edu

Keywords: Google Street View, Pedestrian count, Walkability
Introduction

There is a heightened and widespread interest in building walkable and healthy communities in recent years among researchers and practitioners. Findings from recent research have suggested a link between the built environment and physical activity. Streets, as one important element of the built environment, should be designed not only to enhance mobility choices but also to reinforce walkability, livability, and sustainability (Yin, 2013). Responding to the growing demand for walkable and transit-oriented development, many studies are in an effort to improve the pedestrian environment (Ewing and Bartholomew, 2011). Pedestrian count is a quantitative measure of pedestrian volume to help evaluate walkability and how it correlates with land use, and other built environment characteristics. The count data can also be used as baseline data to help inform planning and funding decisions.

There is, however, insufficient and inadequate research on pedestrian volume and movement even though pedestrian has been the subject of increasing attention among planners, engineers and public health officials. Collection of detailed information about non-motorized activity has been insufficient and inefficient in many transportation and built environment studies. Pedestrian count data has be collected by field work, self-reported survey, or automated counting. Field work and self-reported surveys are more subjective than automatic counts using video-taping or sensors. Most pedestrian counts are done manually because of the high cost associated with using technologies such as laser scanners and infrared counters. Automated counting technology for pedestrian is less developed even though it has been used for many years for motor vehicles.

With the recent rapid development of internet and cloud computing, we are entering the era of ‘big data’ with the “Internet of People and Things” and the “Internet of Everything” (O'Leary, 2013). ‘Big data’ was defined to account for the effort to make the rapidly expanding amount of digital information analyzable and “the actual use of that data as a means to improve productivity, generate and facilitate innovation and improve decision making” (O'Leary, 2013). Video-based and image-based human detection have had a wide range of applications in robotics, intelligent transportation and other fields for collision prediction, driver assistance, and demographic recognition etc. (Prioletti et al., 2013; Gallagher and Chen, 2009). Google Street
View provides panoramic views along streets for many streets of the U.S. and around the world. It is readily available to anyone with access to internet. Even though there are some privacy concerns about these images and some parts of the images such as people’s faces or automobile license plates were blurred, they are still potentially useful to help to identify number of pedestrians on a particular street objectively to generate patterns of walkability across a city. This study explores extracting pedestrian count data using Google Street View images, aiming to provide and recommend future research an alternative method to collect pedestrian counts more consistently and subjectively and stimulate discussion of the use of big data for planning and design.

**Pedestrian Detection**

The current mainstream pedestrian detection algorithms are mainly based on statistical classification method that classifies features of various parts of a human body, followed by pattern recognition methods that are used to identify pedestrian. One of the most influential pedestrian detection algorithm proposed by Dalal and Triggs (2005) was characterized by the histogram based on the gradient direction (Histogram of Oriented Gradient, referred to as HOG) (Dalal and Triggs, 2005). HOG describes the distribution of the intensity and direction of the gradient of the local image region. Based on these distribution features the object can be well characterized with reasonably good detection performance. Continuous improvements have been put forward since then. A typical example is Zhu et al. (2006) that introduced Boosted Cascade Face Detection algorithm to the field of pedestrian detection (Zhu, et.al, 2006). Using this algorithm, the detection rate has greatly improved. As the numbers of HOG features are increased, however, more time for training is needed. Other improvements include MultiFtr and Hog-Lbp that extracted multi-features to improve the performance of pedestrian detection algorithm for environmental robustness (Wojek and Schiele, 2008; Wang et al, 2009), and FPDW that improved image pyramid based on the multi-features to achieve a more rapid and robust pedestrian detection (Dollár et.al, 2010). However, these algorithms cannot overcome the effects of partial occlusion. Felzenszwalb et al. (2008) proposed the Deformable Part Model (referred to as DPM), which is a two-layer model with the global and part characteristics that can overcome the partial occlusion to some extent. On the basis of this model, Felzenszwalb et al. (2010)
proposed a Cascade Deformable Part Model (referred to CDPM), which helped to reduce the detection time (Felzenszwalb et al., 2010).

However, pedestrian detection is still challenging when two or more pedestrians mutually occlude with different overlapping degree. In addition, the efficiency needs to be improved based on the current methods. This paper proposes a new pedestrian detection algorithm based on sparse multi-scale image segmentation and cascade deformable model to develop an automatic method to detect relative positions of pedestrians in Google Street View images to get pedestrian counts. This pedestrian detection algorithm has three phases: model training, sparse multi-scale segmentation and fine detection to get robustness and efficiency through training and segmentation.

The fast pedestrian detection framework is illustrated in Figure 1. First, because of mutual human body occlusions, an occlusion model is trained with latent support vector machine (LSVM) to realize the pedestrian model training offline (Felzenszwalb et al., 2008). Second, a sparse multi-scale image segmentation is designed to extract the regions of possible pedestrians, to narrow detection range, and to eliminate a large number of disturbing areas or background regions to achieve the primary detection. Finally, the Cascade deformable part model is integrated in the region of interest to realize finely divided regions for multi-scale fine pedestrian detection. The occlusion model is initialized with different degrees of overlapping such as 15%-25%, 25%-55%, and 55%-85% etc, and then the features of occlusion pedestrians are depicted as variables that are efficiently annotated and different from those of single pedestrian.
Fig. 1 The Framework of Proposed Algorithm

Model Training

Following Felzenszwalb et al. (2010), the LSVM is used in this paper to train the binary classifiers. The LSVM training objective obeys the scoring functions of the following form.

\[ f_\beta(x) = \max_{z \in Z(x)} \beta \cdot \phi(x, mz). \quad (1) \]

Here \( x \) represents a detection window as input; \( \beta \) is a vector of model parameters; and \( mz \) contains the values to latent variables such as part placements. The possible latent values for \( x \) is set in \( Z(x) \). In analogy to classical SVMs, \( \beta \) is trained from labelled samples of \( D = \{ (x_1, y_1), \ldots, (x_n, y_n) \} \), where \( y_i \in \{-1,1\} \), by configuring the following function,

\[ \beta^*(D) = \arg \min_\beta \lambda \|\beta\|^2 + \sum_{i=1}^n \max(0,1 - y_i f_\beta(x_i)). \quad (2) \]

Where \( \max(0,1 - y_i f_\beta(x_i)) \) is the standard hinge loss. \( f_\beta \) becomes linear in \( \beta \), and linear SVMs as a special case of latent SVMs is obtained, when the latent domains \( Z(x_i) \) is restricted to a single choice (Yu and Joachims, 2008).
Primary Detection Based on Sparse Multi-scale Image Segmentation

The specific procedure of proposed sparse multi-scale segmentation algorithm is illustrated in Figure 2, which can be used to establish streamlined image pyramid quickly and accurately. During this procedure, HOG feature of the pyramid is calculated first. Then the streamlined HOG pyramid is built. Finally, primary detection based on multi-scale image segmentation is realized.

As local image entropy has brightness invariance, which can better express the texture features, it is suitable for coarse segmentation pedestrians. In this paper the sparse multiscale entropies are calculated to depict the texture features. Because pedestrian has obvious texture features which are different from those of in the background such as the ground and the sky, we adopt local image entropy to construct a morphological filter and build a streamlined image pyramid.

![Fig. 2 Multi-scale Image Segmentation](image)

Fine Detection Based on Cascade Deformable Part Model

The deformation and positional relationships between the parts and the object as a whole can be calculated by using the deformable part model, which can adapts to detect non-rigid targets such as pedestrians compared to the previous algorithm. At the same time, this model, however, also increases the complexity of the algorithm, which slows the detection efficiency. Thus, this paper integrates the cascade algorithm to reduce the deformation calculations on which parts there are smaller contribution to realize fast and robust pedestrian detection performance.
The proposed ADPM is formally defined by an adjacent root filter $m_F$ and a set of adjacent part models $(m_{P_1}, \ldots, m_{P_n})$, where $m_{P_i} = (m_{F_i}, v_i, s_i, a_i, b_i)$. Here $m \in \{-t, 0, t\}$, $t \in \mathbb{N}^+$. If $m = 0$, $m_F$ and $m_{P_n}$ represent the middle root filter and middle part model respectively. If $m = t$, $m_F$ and $m_{P_t}$ stand for $t$-th right root filter and $t$-th right part model separately. If $m = -t$, $m_F$ and $m_{P_{-t}}$ represent the $t$-th left root filter and $t$-th left part model. Here $m_F$ is an adjacent filter for the $i$-th adjacent part, $v_i$ stands for a two-dimensional vector defining the center for a box of possible positions for adjacent part $i$ relative to the adjacent root position, $s_i$ contains the size of this box, which $a_i$ and $b_i$ are two-dimensional vectors containing coefficients of a quadratic function measuring a score for each possible placement of the $i$-th adjacent part.

$H$ is a HOG pyramid and $mp = (x, y, l)$ presents a cell in the $l$-th level of the pyramid. The vector obtained by concatenating the HOG features in the $w \times h$ subwindow of $H$ with top-left corner at $m_F$ is denoted in $\phi(H, mp, w, h)$. The score of $m_F$ on this detection window is $m_F \cdot \phi(H, mp, w, h)$. Below $\phi(H, mp)$ is used to denote $\phi(H, mp, w, h)$ when the dimensions are clear from context. A placement of a model in a HOG pyramid is given by $mz = (mp_0, \ldots, mp_n)$, where $mp_i = (x_i, y_i, l_i, p_i, s_i)$ is the location of the adjacent root filter when $i = 0$ and the location of the $i$-th adjacent part when $i > 0$. Here, $p_i$ and $s_i$ contain the position and size relationship between current root filters and next adjacent searching window respectively.

The score of a placement is given by the scores of each filter plus a score of the placement of each adjacent part relative to the adjacent root. Here, the middle filter, right filter and left filter are calculated separately. The specific calculation formulation is as below.

$$
\sum_{i=0}^{n} m_{F_i} \cdot \phi(H, mp_i) + \sum_{i=0}^{n} a_i \cdot (\bar{x}_i, \bar{y}_i) + b_i \cdot (\bar{x}_i^2, \bar{y}_i^2) 
$$

(3)

where $(\bar{x}_i, \bar{y}_i) = ((x_i, y_i) - 2(x, y) + v_i) / s_i$ corresponds to location of the $i$-th adjacent part relative to the adjacent root location. Here $\bar{x}_i, \bar{y}_i \in \{-1, 1\}$.

The score $S^*$ of a placement $mz$ is expressed below,

$$
\beta = (m_{F_0}, \ldots, m_{F_n}, a_1, \ldots, a_n, b_n) \\
\psi(H, mz) = (\phi(H, mp_0), \phi(H, mp_1), \ldots, \phi(H, mp_n))
$$
Experiments and Analysis

To prove the robustness of the proposed algorithm, training data set of this paper is INRIA, using LSVM as the training methods. The model results were validated against two sets of data. One is pedestrian counts data collected from 2011 to 2013 in the springs and falls by the graduate urban planning students at the University at Buffalo, The State University of New York. The second data is a walkability index assigned to streets, which was collected from WalkScore, a private company that provides walkability services to promote walkable neighborhoods. Pedestrian count data was collected by counting number of pedestrians on sample blocks scattered in the City of Buffalo at non-intersection locations, in a 15-minute interval. Walk scores were calculated based on distance to the closest amenity such as parks, schools, etc. The Google Street View image data is static shots for every street. Even though these data sets were collected using different methods, the patterns of walkability reflected from these data should be a reasonable reflection of the how streets in Buffalo are really used by pedestrians.

Findings

The results of multi-scale image segmentation are shown in Figure 3, which shows pedestrians can be segmented effectively. At the end of this primary stage, background was removed, some of the environmental interference was reduced, and the effective area for fine detection stage was retained. This helps to reduce the false detection rate and to increase accuracy because unnecessary complex calculations are avoided at the fine detection stage. Thus the detection rate was significantly enhanced.
Figure 4 shows the results of the proposed algorithm. After using sparse multi-scale segmentation to remove the background area, the green rectangles are detected as specific pedestrian area. As can be seen from the figure, pedestrians can be detected by the proposed algorithm. Experimental results show that the detection accuracy is increased compared with the traditional Pedestrian detection algorithms, and the detection speed is improved by 32%.

Figure 5 shows the results of the pedestrian counts based on the pedestrian detection method proposed in this paper (left map) in comparison with the counts got from the field work (middle map).
map) and patterns of walkability from WalkScore (right map). All three maps show the similar patterns of walkability in the City of Buffalo, with the highest concentration of pedestrian in downtown and Elmwood village areas. Google Street View data has less number of pedestrians than the counts from the field work because Google Street View captures only number of pedestrians with one point of time during one static shot while field work usually captures counts for a period of time.

![Pedestrian Counts Comparison](image)

**Fig. 5 Pedestrian Counts Comparison:**

Google Street View Image Pedestrian Detection vs. field work vs. WalkScore

**Discussion and Conclusion**

This paper proposed a pedestrian detection algorithm based on sparse multi-scale image segmentation and cascade deformable model to extract pedestrian counts from the Google Street View images. The experimental results showed that this algorithm performs 32% faster than the traditional DPM method. The detection results were compared and showed to resemble the pedestrian counts collected by field work. The patterns of walkability in the City also resemble the one from WalkScore data. Future work includes further pedestrian characteristics analysis, combined with pedestrian tracking algorithm to accelerate the detection efficiency, and robust
real-time pedestrian detection.

Current Google Street View images as it is published and made available to the public have two aspects of limitations as follow for pedestrian detection:

1) Metadata. Metadata such as street-view image collection time is not complete and consistent. For example, some of the image collection time is accurate to a specific day, while others are only accurate to month. Such data incompleteness and inconsistency made it difficult to match information extracted from the images with the weather data to do more research on how whether influence pedestrian volume. The U.S. city wide weather forecasts have currently been accurate to the granularity of hours; therefore, if acquisition time granularity of street-view can be published accurate to hours, the validity of the data for research use will be improved.

2) Un-measurable images. The Google Street View data are images acquired by vehicle platform. Street view images are acquired with CCD sensor from different viewpoints. Because images of different angles have different geometric distortions, it results in difficulties to be registered with the vector data. As Google has not provided measurable parameters of the street-view currently, these images cannot be used for geometric measurements directly. If the internal and external orientation elements of the image are released in the future, the stereo image pair can be constructed. In that case, pedestrians can be distinguished from the poles, trash, roadside advertising board more accurately and effectively, and the pedestrian detection accuracy will be higher. Measureable images can also be used to conducted 3D-GIS based calculation and analysis, such as proportion of sky for the research on walkability (Yin, 2014).
References:


Submission for BDUIC 2014 Workshop.

**On City Dashboards and Data Stores**

Oliver O’Brien*, Michael Batty, Steven Gray, James Cheshire, Andrew Hudson-Smith  
The Bartlett Centre for Advanced Spatial Analysis (CASA)  
University College London  
Gower Street, London WC1E 6BT  
* corresponding author. Email o.obrien@ucl.ac.uk
On City Dashboards and Data Stores

Abstract
CASA has become a hub for experimenting and implementing new technologies for simulating and visualising London’s rapidly expanding datasets. This is part of the wider domain of our focus on urban informatics that frames our analysis here. This paper both reviews recent tangible outputs from our laboratory and notes work in progress, some of which, at the time of writing, is just beginning. We illustrate our thesis with many applications, largely applied to large cities, in particular London.

Keywords: data portal, data store, dashboard, urban metrics, open data

1 Introduction

1.1 The Urban Informatics Field – To Inform, Analyse and Educate
Urban informatics has been defined as the use of data to better understand how cities work (CUSP 2014). Of course this is little different from the practice of urban research and practice since urban planning became established institutionally over 100 years ago. Planning has since then been based on the notion first popularized by Patrick Geddes (1915) who argued that survey is essential to our understanding of cities with data pertaining to the city as a prerequisite to effective planning and policy making. In fact, this generic definition of urban informatics as originating in data needs to be qualified in that it is data that is digital that is of particular concern with the data being of many types from ‘small’ to ‘big’ in volume and scale, and from data generated in real time from sensors to data captured from individuals in their responses to social and economic functions.

In some senses, urban informatics is coincident with the smart cities movement which is focused on developing digital methods and tools for improving the performance of cities in terms of their efficiency and their equity (Batty et al 2005). Data is clearly a part of this in that it is the origin for the development of any intelligence that can be applied and embedded in the city, making it ‘smart’ or at least its citizens ‘smart’. There is nothing in this general notion of the smart city that is particularly digital but this terminology and our focus is here adopts these ideas with respect to the digital world in general and big data in particular. Our examples will illustrate these notions as we continue.
Due to the comparative newness of these ideas, many of the tools that define urban informatics have multiple roles. There is a clear mandate to inform the way in which computation is being embedded in the city. In this sense, the tools that we introduce here which are all framed around the ideas of dashboards and data stores tend to produce information that provides a synoptic picture of what is happening in the city on a relative frequent basis, thus aiding understanding and providing fuel for defining urban problems. Dashboards and data stores imply usage for analytical purposes and coupled to such tools is the whole arsenal of what is now being called urban analytics which consists of the array of tools fashioned to deal with digital data in descriptive and predictive contexts which is key to informing policy makers and planners about urban problems. This analytic role is key to the planning process but in general, these tools also provide ways in which a wide array of groups interested in the city might be educated. As we outline below, pedagogy is an essential feature of new tools fashioned for the current digital age and the uses and applications of these ideas and their data is so new that we need to learn about what is possible through applications: this is education in the widest sense – pedagogy – which is a key feature of all our examples which follow.

1.2 CASA’s Research and the Urban Informatics Agenda

Our research programme has evolved from an original focus on the spatial analysis of urban phenomena to the development of a science of cities through the medium of informatics. Our concern is for simulating how city systems evolve and how they can be steered to meet certain objectives through planning and one of our dominant themes is the development of visual resources to enable us to understand and predict the future of our cities. Our focus on data is key to the various simulation models that we are building and these range across different spatial scales from the level of the local movement of pedestrians and vehicles to the location of activities at the scale of the metropolis. The particular focus of our work on simulation is using various land use transportation interaction (LUTI) models to examine the impact of large infrastructure projects such as airports, new transit systems, housing developments and new locations for employment and retailing and these models traditionally have not required big data. However the emergence of streamed data particularly in social media and transport from mobile and fixed sensors is enriching the data that such simulations address, and our agenda has begun to embrace these developments, shifting the focus of our models away from cross sectional simulations to temporal changes in the evolution of cities.
We are also developing more physically oriented models of the development process using cellular automata models as well as fine scale agent based models of transportation using the MATSIMs platform. All of these models require massive visualization to make sense of their data and processes. Much of this visualization is being developed in terms of informatics, specifically in terms of the analysis of big data (Serras et al 2014), and the examples we will explore here are basic to this wider agenda. In particular, a series of projects in CASA have focussed our visualisation efforts on the translation of spatial data to the third dimension and our Virtual London model which is a 3-D virtual city model based on extruded land parcels, laid on a digital elevation model where building or rather block heights are determined from LIDAR data has been used to fashion various visualisation such as our ‘PigeonSim’ fly-throughs (Batty 2014). Our London data table and various physical realisations of visual flows through projection on the table are part of a wider effort in CASA to disseminate digital media through more tangible artifacts [5]. Our work with fashioning real time data in an accessible visual form as dashboards is part of this effort.

1.3 Urban Data Theory – Official, Social, Sensor

Each of the numerous sources of data relating to the urban realm can generally be categorised as one of three types – official, social and sensor. These types typically differ significantly from each other in terms of authority, frequency, timeliness, detail, distribution and quality, but together they can give the observer a rich observation of a city from multiple perspectives.

Official data sources have been the traditional structural blocks of urban informatics displays. These are data supplied by local or national authorities under an official capacity. Examples include population estimates, crime rates and measured delays on public transport. These datasets are by definition authoritative and can be regarded with trust as they have been vetted and verified by the publishing organization prior to release. However the timetable of release can typically be quite long and frequency low – with some notable exceptions detailed later – typically being published on a cycle of months or years, and often some time after the observed event or measuring period. Census datasets are a good example of this type of data – often serving as the only definitive measure of many metrics but typically only appearing every 10 years (in the UK) but with intermediate releases containing estimates to the changes in some key measures. Some authorities – particularly transport ones – have
recognized the benefit of publishing rapidly changing information in an official capacity. Transport for London, the city’s transport authority for most public transport services, through its Open Data Users portal\(^1\) provides almost real-time information on availability of buses (TfL 2014a) and metro (TfL 2014b) services at bus stops and stations respectively. This has allowed an ecosystem of third-party application development to flourish, with sophisticated examples such as CityMapper\(^2\) complementing the authority’s own consumer website. Format availability also varies greatly, from simple websites and PDF reports to JSON/XML data feeds, in the case of transport data, sometimes cloud-hosted and able to scale to manage demand spikes during periods of disruption.

Complementing and yet starkly different to the official sources, social media data – principally from Twitter as the other major platforms popular in most countries in the western world (Facebook and Google Plus) are more difficult to obtain information from - can provide huge volumes of information, directly from people “on the ground”, on a very timely and rapidly updating, but with highly variable quality and accuracy. Individual tweets in general can be poor in quality. However, aggregating individual messages can still be a useful indication of a city-affecting event, such as a significant public transport disruption. Text-mining to detect keywords, geolocation information, and sentiment analysis, when applied to multiple “tweets” can reveal information. It can outperform even fast-acting official data feeds in certain cases. A study by Aguilera (2012) looked at disruption on the RER metro network in Paris, comparing a social datasource (Twitter) with official data (the operator website) and sensor data (mobile phone cell-tower connection changes) to observe the effectiveness of each on detecting and measuring the disruption. Examples of social media data types include simple messages from anyone, messages from nominated authoritative accounts, trends in keywords and volume spikes across an identified area.

Finally, sensor data can be used to measure natural phenomena, to complement the other data sources. This information is normally accurate, as long as devices are calibrated and maintained. Many newer sensor devices come with internet connectivity so can broadcast their readings in in near-real time. Both official sources (e.g. traffic counters installed by local authorities) and personal sensors (becoming rapidly popular with the quantified self

\(^1\) [https://www.tfl.gov.uk/info-for/open-data-users/](https://www.tfl.gov.uk/info-for/open-data-users/)
\(^2\) [https://citymapper.com/london/superrouter/](https://citymapper.com/london/superrouter/)
movement\(^3\) can output rapidly updating information with rigid taxonomies and useful metadata. Typical useful sensors in an urban context include weather observations and measurements, air quality recording and vehicle counters for congestion/popularity metrics.

## 2 The CityDashboard Project

### 2.1 CityDashboard as a Website and Collection Service

CityDashboard is a website, collection service and API that consumes and redisplay urban informatics for a number of cities in the United Kingdom. It combines official, social and sensor data into a single page view. One key principle in its design is that the information it displays is rapidly updating. As a rule, all information shown on the site should update at least once a day, and preferably at least once an hour during the working week, with a few specific exceptions.

![Image of CityDashboard website](https://example.com/citydashboard.png)

**Figure 1** – The CityDashboard website, showing urban informatics for London.

---

\(^3\) [http://quantifiedself.com/](http://quantifiedself.com/)
CityDashboard is designed in a modular fashion. Each module processes and represents one external datasource. Data is requested using Python scripts. Common functions are used to retrieve the data from a specified URL and cache the result for a specified period to prevent unnecessary requests to external servers. Other standard functions are used to derive a status colour related to the information, using a specified colour ramp and low/high range.

Modules are configured to output in one or more of three formats, HTML, CSV or grid. The HTML is used by the CityDashboard website itself (Figure 1). HTML is output in a rectangular box, typically with a coloured background and the number, or numbers, of the metric concerned, shown prominently. CSV format is used by the “Map” view on the CityDashboard website (Figure 2), which uses point latitude/longitude coordinates contained within the data, where possible, and also is the format for the API, used by some external services. Finally, the grid view outputs HTML but consisting simply of coloured boxes (Figure 3) – designed to be integrated into a simple artistic view of a city’s current data footprint.
Figure 2 – The map view for CityDashboard. Data items are represented by coloured points on the map. The highlighted point is shown with a thick purple border and its content is shown on the bottom right corner – in this case, a traffic camera live view.

Figure 3 – The grid view on the CityDashboard website, for London, with a mouseover tooltip on the “air quality” module.

The CityDashboard website has a straightforward look – all data for a city is presented on a single screen, with colour hues and lightness variations used for the data elements themselves rather than for the background interface, which is predominately greyscale. Each module is included in a panel, retrieved via the HTML format described above. An attribution link to
the source website is always included, along with a countdown timer that indicates when the panel will next be refreshed with new data. These timers change every second, giving the website a constantly evolving appearance even if the data itself is not currently changing. Figure 4 shows an example of a module, with title, attribution, timer and the visualised data itself.

![Figure 4](image)

**Figure 4** – A module’s HTML output, as displayed on the CityDashboard website. The attribution link is bracketed. The number at the top-right is the countdown timer, highlighted in orange to show a refresh appearing in the next second. The green background colours indicate that the numbers displayed are within normal bounds.

CityDashboard is configured via text files and making a new city available on CityDashboard is as simple as adding a text file for that city, which defines the modules to be used, the module display order and update frequencies, and the city-specific IDs and URLs for the underlying remote data sources. The service is currently configured for eight British cities, as well as a special view for UCL Museums. This uses a number of the London modules, tuned for the UCL locality within the city, as well as bespoke modules displaying specific museum information, such as current opening status, room monitoring data and a sample item from each museum collection.

The website contrasts with other popular urban dashboards, which are discussed later, by not making use of digital representations of dials, meters or other traditional monitoring apparatus – instead the data itself is the predominant feature displayed. While dials and other visual effects can be used to catch the eye to indicate data changes, they can be sophisticated to set up and can reduce the quality of information visible in a single screen.

### 2.2 CityDashboard as an API

It is acknowledged that the CityDashboard website is relatively simple looking and has a minimalist design. The project was designed first and foremost as a platform, for use by the
accompanying website as a conceptual demonstration but also by other services, both internal to the project and external. As such, the project has an Application Programming Interface (API) and the processed and formatted data is made available in a consistent style. The API is roughly documented on a website\(^4\) accompanying CityDashboard.

The format is CSV (comma-separated values) with different files for each module within each city. The data is accessed via a simple URL call with the appropriate parameters for the city and module included, for example \(\text{http://citydashboard.org/modules/osmchangesets.php?city=\textit{london}&format=csv}\) – in this case, this supplies data from the OpenStreetMap changesets module, for London, in the CSV format. This particular module retrieves lists the most recent edits made to the OpenStreetMap project, for map features in the London area.

A header row contains metadata, such as attribution links, the module’s theme colour, refresh rate for that module/city. This is then followed by a row containing column headings, and then a row for one or more data points for each module. Each data point has an associated location (which can be empty for modules not designed to be shown on a map) as well as a status colour, and the information itself, which may be a simple number (e.g. a share price), a text string (e.g. a tweet) or multiple items (e.g. air quality rating, measured value and measurement type).

It is acknowledged that the CSV format, while simple to implement and very human-readable, is a poor format for a flexible API like CityDashboard’s. It lacks the self-describing capabilities of more sophisticated formats such as JSON and XML and its necessarily a “flat” structure rather than being hierarchical. The use of JSON, including GeoJSON for describing spatial data, is an emerging standard for APIs and most likely one that will become relevant in making data from future dashboard-like projects open.

### 2.3 API Use Examples: The London Periodic Table and Prism

A number of projects use the CityDashboard API. These include the London Periodic Table\(^5\), which was created as a more vivid, bespoke version of the CityDashboard website, focusing in a single location in London, specifically the locality surrounding the CASA office. All data in the London Periodic Table is presented in a square, with a single colour, a primary value, and a minimum of supporting information, such as a caption, unit or secondary value. The

\(^4\)http://oobrien.com/2012/06/citydashboard-the-api/

\(^5\)http://casa.oobrien.com/periodictable/
colour in each square pulsates if the value is unusually low or high. For example, the current observed temperature from the CASA weather station shows in a panel, with a colour ramp from blue (cool) to red (hot). Temperatures above 30°C, unusual for the central London location that the London Periodic Table is focused on, cause the square to pulse from dark red, to bright red, every few seconds. Since the initial creation of the visualisation, the collection of squares that form the visualisation have been augmented by an additional column of squares for indoor measurements (Figure 5).

Figure 5 – The London Periodic Table, using the CityDashboard API. Indoor measurements are shown in a separate column on the right.

Prism (Matsuda 2012) was a sculpture created by digital/new media artist Keiichi Matsuda and workshop assistants. It consisted of a 3D object made of a series of connected triangles. A number of projectors, internal to the structure, shone onto each triangular face a visualisation that received data from the CityDashboard API, transforming it into a texture, a sequence of words, or another visual effect. The sculpture was suspended from the ceiling of a gallery in the Victoria and Albert Museum and also protruded through the floor into another gallery below (Figure 6). A nearby staircase allowed visitors to look over the city itself, and contrast their view with the data view presented on Prism. The sculpture was a temporary work, although an accompanying website remains, with the textures that were projected remaining visible.
Figure 6 – The Prism sculpture, hanging in the Victoria & Albert Museum. Each of the visible panels is showing an internally-projected visualisation of data from the CityDashboard API.

3 Informing the Urban Population

3.1 Urban Data Stores

Urban data stores have become a popular way for more enlightened and forward-thinking city authorities to share their public service information as open data with their population in a transparent and straightforward way, typically through a website showing an index and metadata, and then either an API or downloadable files, sometimes as a CSV but often in less accessible formats such as Excel or PDF files. Berners-Lee (2006) has proposed a five-star system for rating such data, with the availability at all gaining a single star, up to five stars for linked self-describing data that is in a uniform location and is in a machine readable and open format.

Different cities release different types of data. The data can vary from static information such as locations of facilities, to near-real-time feeds such as the live running of public transport services. Data stores serve to collect and aggregate the data, although in some cases they exist as catalogues or directories, rather than containing the data themselves. This can be due to the diverse nature of the nature represented, and the many public authorities in charge of each class of data, and a desire that the original publishing entity remains the canonical source of the data.
The potential audience for a data store is varied – it could be journalists looking to create a story out of data, developers creating a smartphone app to help people search for their nearest facilities, academic researchers looking to understand city patterns or develop and test transport models.

3.2 Some Examples of Urban Data Stores

Chicago’s Data Portal\(^6\) is one of the oldest and most comprehensive city data stores, covering a huge range of topics from energy to crime, expenses reporting and the locations of farmers’ markets. It is also straightforward to obtain geographic boundary files, for users mapping the datasets or carrying spatial analysis of them. The focus is very much on the data itself, with the catalogue list and data displays dominating the design of the online portal. Boston has a very similar datastore, using the same underlying web software as Chicago’s.

The London Datastore\(^7\) was launched in 2010 by the Greater London Authority, with the personal backing of the city’s high profile mayor. London’s governing structure is complicated, with much of the day-to-day running controlled by the 33 boroughs, transport functions devolved to Transport for London which has its own open data portal, and other functions handled by the national government which also had a data store. As such, it was a challenge to create the platform. Coleman (2014), who architected the store, found the different degrees of willingness to be open, between the various bodies concerned, to be a challenge. The London Datastore acts mainly as a catalogue of data, with a limited API available but little data actually contained within the store itself. As such, it has proved difficult for the catalogue to remain up to date, as source URLs have moved or updates have been published at source, without notifications filtering through the to the London Datastore itself.

Paris\(^8\) launched a data store shortly after London, however it is smaller in scope with around 100 datasets available on topics as diverse as the most popular books loaned in its libraries, to the availability of bikes in its bikesharing system and locations of wifi hotspots.

\(^6\)https://data.cityofchicago.org/
\(^7\)http://data.london.gov.uk/
\(^8\)http://opendata.paris.fr/page/home/
Washington DC has long had good data accessibility, stemming from the accountability and scrutiny that would be expected in a city containing the national seats of government. Its Open Data Catalog\(^9\) again focuses on making the data as easy to download as possible, with many datasets visible on the front page of the website, complete with one-click abilities to view the data on a map.

### 3.3 Official Data Dashboards from City Authorities

City authorities sometimes build data dashboards, often to showcase some of the outputs from their data stores. These are generally very different in concept to CityDashboard, despite showing the “single page” look and having a focus on headline numbers, the content is from official data sources only. In London’s case\(^10\), the somewhat arbitrary selection of data metrics shown, can make it seem that the data shown is cherry-picked to present the city in the best light. Many of the statistics shown only update on a monthly or even yearly basis, and it is questionable the value of seeing a set of red/green arrows (Figure 7) at a glance in terms of policy decisions or an overall feel for how the city is doing, right now.

![Figure 7 – The London Dashboard.](image)

Washington DC has a number of official dashboards. Its Green Dashboard\(^11\) is, like London’s extremely simple general, rating just six metrics by one of three categories (Figure 8).

---


\(^10\) [http://data.london.gov.uk/london-dashboard](http://data.london.gov.uk/london-dashboard)

\(^11\) [https://greendashboard.dc.gov/](https://greendashboard.dc.gov/)
In Amsterdam, WAAG, a technology-focused collective, recently created a City Dashboard for the city\textsuperscript{12}. It is particularly notable for its chart display – using graphs to show the recent trends in the reported statistics, as well as the current values, and estimated short-term future readings. For example, the diurnal variations in air quality measurements, are clearly displayed (Figure 9). More static information such as population density is also included. Because WAAG is not an official dashboard for the city, they are able to straightforward include social media data (relating to the political parties and football teams linked to the city) and sensor information, too. The Dashboard has been very attractively designed, while allowing the data to be the dominant visual feature.

\textsuperscript{12} http://citydashboard.waag.org/
3.4 The OKFN Open Data Census

In a commendable effort to pull together the various open data outputs of cities around the world, the OKFN (Open Knowledge Foundation) has created a census of city data outputs. The census is user-contributed and editor-verified, and ranks and rates cities on a number of tests for different kinds of datasets. The project is being done on a country-by-country basis to reflect the different data cultures at a national level – some countries retaining certain data types at a national rather than city level. The census acts to rate the cities but is itself also a useful means to discover the datasets in the first place. The census currently covers cities in 25 countries.

Such a rich global repository of open data sources for cities could potentially lead itself to a powerful and impressive set of dashboards covering many cities – possibly even a dashboard of dashboards showing how each city is moving towards becoming an open and accountable environment.

Figure 10 – Part of the OKFN Open Data Census website for US cities. There are 18 categories, and each category is scored Yes/No/Unsure for 10 metrics, such as whether the data is available for free, and whether it is updated on a timely basis. Each city is then scored based on the results – currently San Francisco scores most highly.

4 Visualising the City – the Next Level

http://meta.census.okfn.org/
4.1 The London Data Table

Despite the age of satellite navigation systems, slippy maps on smartphones, and other new mapping formats, the paper map remains a popular way to navigate. It allows the user full control of where they want to look, and its tangible feel is familiar. With this in mind, CASA was keen to exploit the CityDashboard API and other urban-informatics projects and visualisations created in the lab, onto a single, physical product. The London Data Table is a large (1.5m x 2m) wooden table. A programmable lathe was used to cut the outline of the Greater London Authority boundary. While this is not entirely conterminous with London’s urban data extent, it still represents a well defined boundary for London as a city, and is one of several popular definitions of the metropolis – many of the informatics datasets for London, particularly official ones – make use of the boundary and the 33 boroughs (local authorities) it contains. The table is painted white to act as a screen, and a short-throw projector is mounted directly above the table, attached to it via a pole. A small computer was bolted to the projector, and a script loads animations created in the Processing Java IDE, or prerendered film-clips. In both cases, the content is output to a specific projection (British National Grid), resolution (1024x768 pixels) and in full-screen. This allows the content to line up perfectly with the physical table (Figure 11). Basic information about the current visualisation shown, including a key and original source information, is included on an iPad mounted below the projector.

![Figure 11 - The London Data Table, here showing data from the CityDashboard API.](image-url)
The London Data Table cycles between a view of London’s informatics data using the CityDashboard API, an animation of recent activity in the city’s large and active bicycle sharing system, aircraft positions and speeds detected by a flight radar, and videos of events in London which had a significant spatial element – namely, bombsites from a night of the London Blitz in 1940, and the locations of rioters – and rioters’ home addresses – from the 2011 London riots. Finally, it also includes a video of timetabled public transport movements (buses, coaches trains, trams, metro and river-buses) for a typical weekday. This could potentially be visualised on an as-live basis but the computation power needed to do this is significant, due to the number of datapoints and London’s complicated networks, so a pre-rendered video was chosen.

The table has been demonstrated at a number of fairs and exhibitions and the novelty of viewing urban informatics by leaning over a flat table has proved a compelling experience. It could perhaps be compared to the stereotypical “war planning” table where resources are positioned by hand, the ability to view the area (or city) from any angle allowing for a different perspective.

4.2 PigeonSim

PigeonSim is a simulation where users “fly through” a 3D Google Earth environment of a city, such as London. A Microsoft Kinect device detects motion and maps it to the projected movement. Data from CityDashboard is fed into the simulation, using the KML file format, where its appears as a number of pins placed through the city. Users can fly to a pin to see the information, such as the current air quality readings or a message displayed on electronic dot-matrix traffic information displays, that appear on the real-life roads. It was developed at CASA by George MacKerron.

4.3 Tweet-o-Meter

Social Media is arguably one of the most prevalent topics facing academics and policy-makers today due to the recent rise in popularity of social networks over the last decade. Social media is the largest, richest, and most dynamic evidence base for human behaviour, and brings new opportunities to understand groups, movements and society. The Tweet-o-Meter[^14] is a visualisation of the vast number of Twitter messages, or Tweets, generated by

[^14]: [http://www.casa.ucl.ac.uk/tom](http://www.casa.ucl.ac.uk/tom)
public users from 16 major cities around the world. This visualisation consists of 16 digital
dials, which show the number of tweets collected for each geographical local on a scale
between 0 and 1500. The upper value of each dial is due to the number of tweets that can be
collected from the Twitter API in a period of 1 minute before the system collecting the tweets
is blocked from collecting data.

These dials are updated every second to give a real time view of each of the 16 cities around
the world: New York, London, Paris, Munich, San Francisco, Barcelona, Oslo, Tokyo,
Toronto, Rome, Moscow, Sydney, Hong Kong, New Delhi, Shanghai and Sao Paulo. Each
dial, or gauge, displays tweet per minutes (TPM), an artificially constructed value calculated
from the aggregate number of tweets collected by the system for a 30km geographical region
around the administrative centre for each of the host cities. The Tweet-o-Meter system,
which serves as a collection platform for other services inside CASA, is designed to mine
data for later analysis relating to furthering our understanding of social and temporal
dynamics for e-Social Science within the Twitter demographic.

A physical version of Tweet-o-Meter consisting of 12 ammeter dials mounted on a custom
designed backing board was created for the Growing Knowledge’ exhibition held at the
British Library held from October 2011 to July. This Digital to Analogue representation of
user generated social media data has been used to engage the public and provides a talking
point about the challenges of collecting and analysing large sets of data in real-time.

4.4 The iCity Platform

The iCity project, funded by the European Union (Fiore 2012) and cosponsored by the
Greater London Authority (GLA), is an active and ongoing research topic at CASA. It aims
to open up existing infrastructure owned and managed by city authorities, to the developer
community, to create applications for the public good. The project currently has four
European cities involved. Its main output is a platform, providing a single API to access
from, and potentially contact, the appropriate infrastructure available in each city. In
London’s case, which is CASA’s principal involvement, this includes the air quality sensor
network (AQSN) run by King’s College London and traffic volume detectors. The other
cities are variously moving their weather, public transport information and other data sources
to the platform.
The specification and quality of the API is a key consideration in the project. This will likely allow and encourage an ecosystem of applications that can consume and transmit data on the platform, will develop. There are tentative plans for the London Data Table, and CityDashboard in general, to adopt the API from iCity as one of its data sources, rather than parsing the information direct from the underlying source, and protecting it from format changes.

5 Big City – Big Data

5.1 BODMAS and DataShine

BODMAS (Big, Open Data Mining & Synthesis) is a research project at CASA focusing on managing and visualising big data, particularly open data, such as that made available by London’s transit authority, Transport for London. Its initial output has been the DataShine mapping platform, which is currently being used to show another of large datasets relating to the United Kingdom, such as demographics identified from aggregate tables of the most recent census, and government-calculated poverty measures such as the Index of Multiple Deprivation. The platform is likely to become useful as more sophisticated way of showing urban informatics data, such as those currently shown in the CityDashboard map or projected onto the London Data Table.

While BODMAS and DataShine are designed to be processors and viewers of general datasets, rather than those relating specifically to urban informatics, their use of maps as the central way to display relatively large volumes of data, works well in a city context.

5.2 Expanding CityDashboard – City Trends

CityDashboard has always been an “at a glance” view of city informatics. However, in an attempt to capture the time-evolution of the data – diurnal, weekly and season fluctuations, the key information from some of the modules has started to be recorded into a database table. The flat structure of the CSV API has enabled this to be done relatively easily, with each sensor/module/city/time measurement recorded as a single row in the database. The key value is captured, along with any secondary measurement and associated colour. Despite the structure of the API, the data, being from all three data categories (social, official and sensor) is still very heterogeneous and so recording the data results in a very unnormalised table of results, with many empty cells. However it should still be relatively easy to query the table
and retrieve similar information across multiple modules and/or cities, for further visualisation in a chart or other package.

6 Concluding Discussion

In this paper we have explored a number of ways to visualise urban informatics, looking both at official methods (through data stores and official dashboards), and through CASA’s own approaches such as CityDashboard and its associated API uses. We have also discussed more unusual methods, differing from looking at a computer/TV screen, such as the feeding of urban data into a landscape, to be explored by flying around it, or by moving around a projection on a table (the London Data Table). We have also looked ahead to including trend information on CityDashboard, by storing the most recently collected data for graphing, as has already been done by the notable WAAG City Dashboard for Amsterdam.

Urban informatics is still a newly emerging field. It is one which is seeing many ideas and a number of prototype implementations coming to fruition – but one where the is a considerable research and visualisation potential to be exploited. The great rise of the so called ‘big data’ adds challenges of its own – not least in its sheer size – but also increases the richness, frequency and quality of data available to be visualised and explored when trying to understand our cities, inform decision makers and make them better places to experience and live in.

7 References


Centralized Real-Time Platform as a Tool for Urban Studies

Orlando Cabañas
Politecnico di Milano
orlando_cabanas@live.com

Céline Steer
Freelance consultant
celine_steer@hotmail.com

This paper is an attempt to analyze the collection and use of real-time data in the urban space through a centralized platform. First, it seeks to identify the process in which data is collected in an attempt to understand how it has evolved to its current point. The second part presents the real-time data platform and explores the initiatives carried out so far to consolidate all of this data in one single platform, while the last part of this paper focuses on the advantages limitations of having access to all this data together in one centralized place.

Key words: Real-Time Data, Centralized Data Platform, Big Data
**Table of Contents**

I. Introduction ........................................................................................................................... 1

II. Where does data come from? ............................................................................................... 1
   From data storage to data analysis ...................................................................................... 1
   New trends for data collection (from small data to Big Data) ............................................. 2
   From static data to dynamic data (Real-time data) ............................................................. 2

III. From Fragmented to Centralized data ........................................................................... 3
   Structured and unstructured data ....................................................................................... 3
   Big Data Input .................................................................................................................... 4
   Big Data Storage ................................................................................................................ 4
   Big Data input/output processing ...................................................................................... 5

IV. Urban Layers and Data ..................................................................................................... 6
   Demography ....................................................................................................................... 6
   Travel and transport ........................................................................................................... 6
   Natural environment ......................................................................................................... 7
   Built infrastructure ............................................................................................................ 7
   Information society ........................................................................................................... 7
   Economic aspects .............................................................................................................. 8
   Social aspects .................................................................................................................... 8

V. Urban informatics and the emerging science of cities ....................................................... 8

VI. Case studies ..................................................................................................................... 9
   Operations Center in Rio de Janeiro Brazil .......................................................................... 9
   Open Data for cities .......................................................................................................... 10
   London dashboard ............................................................................................................ 12
   WikiCity (MIT) ................................................................................................................ 14

VII. Critical Analysis of Centralized Platform .................................................................... 15
   Benefits ............................................................................................................................. 15
   Cross-field analysis ........................................................................................................... 16
   Real-time information ....................................................................................................... 16
   Visualization ..................................................................................................................... 16
   Decision-making ............................................................................................................. 16
   User-friendly .................................................................................................................. 17
   Model creation ............................................................................................................... 17
   Limitations ....................................................................................................................... 17

VIII. Concluding remarks .................................................................................................... 18

IX. Bibliography ................................................................................................................... 19
I. Introduction

With rapid advances in computer power and data analysis, new ways of representing intricate patterns of cities are emerging. Tools providing real-time and large amounts of data are contributing to making more effective and sustainable decisions for cities, adapted to the rapidly-changing behaviors of their inhabitants.

This paper will analyze the collection and use of real-time data in the urban space through a centralized platform. First, it seeks to identify the process in which data is collected in an attempt to understand how it has evolved to its current point. The second part presents the real-time data platform and explores the initiatives carried out so far to consolidate all of this data in one single platform, while the last part of this paper focuses on the advantages limitations of having access to all this data together in one centralized place.

II. Where does data come from?

Data, hereon used as a term to describe bits of information, is indeed nothing new. However, for most of human history, data was scarce, relatively hard to get, time consuming and resource intensive. Furthermore, the use of this data did not achieve its real potential not only due to data scarcity but because of a lack of knowledge and technological tools that also led to data misinterpretation: as Miller states “one can always search long enough and find a complex but often spurious model that will fit a dataset arbitrarily well” (Smyth, 2000).

From data storage to data analysis

Large amounts of data, collected in datasets, were possible due to technological advances and economic efficiencies in sensors, digital memory, and data management techniques in the 1990s. As described by Robert Kitchin (2013), examples of very large data sets consists of national censuses and government records that provide information about cities and their citizens. Businesses also generate significant amounts of data about their operations, markets and customers.

Whereas in the early days of databases, these were mostly used to store, access and manage data, data owners had an increasing desire to analyze this data. But the complex nature and size of datasets meant that this was only possible if computer scientists carried out data analysis through the application of algorithms (Smyth, 2000).
New trends for data collection (from small data to Big Data)

With unstoppable technological advances, the world has seen an exponential increase in the amount of data stored. As Smolan and Erwitt point out, "from the beginning of recorded time until 2003, we created 5 billion gigabytes of data. In 2011 the same amount was created every two days." (Smolan and Erwitt, 2012). Another source, Zikopoulos (2012), predicts that by 2020, the amount of data stored will reach 35 zettabytes (ZB).

Recently, this phenomena has been coined as Big Data, and although there is not a set academic definition, some of its characteristics is that it is:

- Massive in volume, as described in the previous paragraph
- High velocity, in the sense that it is obtained almost immediately,
- Varied, as it is recorded in different times and locations
- Extensive because it embodies a larger sample

As noted previously, technology and decreasing costs have been a leading factor in the increasing creation of data, and urban data has not been exempted from this. Through automated data, the integration of technology in our everyday lives has facilitated the procurement of data and measurement of flows and interactions related to urban settings. Sensors have been widely used for two main purposes: they have been used as monitoring devices, such as for rainfall, air quality, and traffic among others. Sensors have also been used in everyday transactions, for example to purchase and validate transportation titles, tracking, and identification. Additionally, data is also generated by the communication of and between machines, more commonly known as the internet of things (IoT).

From static data to dynamic data (Real-time data)

Ubiquitous computing, allowing to be constantly connected at any time and place, contributes to the continuous feed of data. This contrasts immensely with the static nature of the data that was formerly collected which did not guarantee accuracy. An example of this are origin/destination (O/D) models, in which surveys are conducted to identify the origin and destination of citizens in order to create plan transportation methods and infrastructure more efficiently. There are two main problems with this form of data collection. The first one is that it is very limiting in the sample that can be obtained. The second limitation is that knowing the
origin and destination of an individual does not allow the surveyor to know the intermediary stops of the person.

With new technologies, such as Global Positioning System (GPS)-enabled devices and data generated by cellphones, used by virtually everyone, it is possible not only to know the exact path of individuals, but also to get this information in real time. This turns every individual and every thing into a potential data producer.

III. From Fragmented to Centralized data

Cities are ecosystems, in which the interaction of different elements creates complex patterns. The large amount of data generated and the exponential trend that seems to be here to stay in terms of data creation increases the probability of not fully exploiting the potential of having valuable data readily available in decision-making processes. Indeed, the high number of sources makes finding and extracting the desired variables difficult. Making sense of the data becomes another problem of its own in addition to being a time-consuming and complicated process.

A potential solution to address this problem is the construction of a cloud-based platform in which all the data useful for urban studies, including real-time data, is centralized in one place, readily available to consult at any time to anyone and can be visualized. The following paragraphs provide an overview of such platform, including the different steps and elements that would compose it.

Structured and unstructured data

Data can be categorized in two types: structured and unstructured. While the former relies on a well-organized system that makes it easy to search for information, the latter is disorganized and hard to categorize. Despite the advantage of having structured information, the truth is that both unstructured and structured data exist and both must be taken into account when integrating data, especially with the increasing use of multimedia and visual material to portray information. Social media, an important element in today’s urban settings is also an example of unstructured data.

Just thinking about the difficulty for a computational system to decipher and categorize content of unstructured data gives an idea of the difference and challenges posed by this type of data when doing data analysis. Fortunately, technological advances have permitted the
development of intelligent softwares to extract important and relevant information of unstructured data.


**Big Data Input**

One of the main characteristics for the proposed platform is that it is open source, meaning that the data content of the platform can come from virtually any source that generates data. As described previously this includes people, organizations and things. For example, if a researcher carried out a study in which he collected data on traffic, he can upload his dataset directly on the platform, following certain parameters. For data coming from the internet of things, such as house energy consumption or traffic sensors, it could be recorded directly if the platform is set as one of the data receptor. Similarly, data coming from social media could also be linked to the platform and retain the commonly used hashtags.

While having parameters are necessary to establish some standards to allow certain functions and comparability, these must remain flexible enough so as to not discourage the researcher to upload the information and the data retriever to access it. The goal is to develop a common language that is user-friendly in order to identify the data through tags, keywords leading to a more convenient and successful search and retrieval of desired data.

**Big Data Storage**

All the data uploaded on the platform needs to be stored in physical computing units. Nonetheless, in order to be functional and adapted to today’s needs, the platform must be linked to a cloud-based system that can guarantee real-time data streams, quick accessibility from anyone and the ability to easily filter it. The cloud-based platform is not only convenient for accessibility purposes (it allows access from anywhere provided there is an internet connection), but it also makes accessing the data a much more speedy and efficient process. Indeed, storing such a large quantity of data -as would be the case of the proposed platform- requires multiple storage locations, which would make accessing the data a slow process because it means managing the data over multiple storage locations. (http://newsoffice.mit.edu/2014/storage-system-for-big-data-dramatically-speeds-access-to-information-0131)
Big Data input/output processing

Data must be processed through a system of code-base software that transforms this data into a user interface. This is valid in the input process in which both structured and unstructured data must reach the platform in a uniform format.

Similarly, the data must be processed for its output so that information can be visualized in an easy to understand manner. Furthermore, the software must enable cross-referencing data variables so that data can be constructively combined in response to a user’s query. (Ratti 2006)

In its conceptual form, the proposed platform resembles the Wolfram|Alpha project, which provides knowledge and answers based on built-in data, algorithms and methods. As stated on their website “[they] aim to collect and curate all objective data; implement every known model, method, and algorithm; and make it possible to compute whatever can be computed about anything.” (http://www.wolframalpha.com/about.html)

Figure 1: Conceptual diagram of integrated platform and functioning
IV. Urban Layers and Data

For the purpose of this paper, seven layers or urban data considered were identified. In this paper, layers is a term used to identify the different categories for the flows and networks proper to the city. The seven layers are: (1) Demography; (2) Travel and Transport; (3) Natural Environment; (4) Built Infrastructure; (5) Information Society; (6) Economic aspects; (7) Social Aspects. Cities are the result of the synergies between layers. Thus, identifying and quantifying these synergies and interaction is a critical exercise to better understand cities and in turn plan accordingly.

Demography

People are the most important elements in urbanized areas and the sole reason cities exist. Quantifying the number of people in a city, their origin, their age, their civil status among others is perhaps the most basic type of urban data and one that has been collected for a long time mostly through censuses and other government data. For city-planners and decision-makers, this is crucial because it gives a snapshot of the city in order to know what type of services are needed, but also to predict for future services.

City governments have been entering also the digitalization era and slowly transitioning and providing digital and record keeping and services, which would facilitate the integration of such data into the proposed platform. However, until the digitalization process is done on a systematic basis, city agencies can become data contributors of the platform by adding the already-existing datasets relating to births, deaths, nationalities, civil status, residency, etc.

Travel and transport

As cities become bigger and an increasing number of people move into urbanized areas, it has become more difficult to see and understand the way people move through its physical space. On a daily basis, thousands and millions of people take on their daily routine and leave behind a footprint whether people move by car, a bike, train, walk etc. This footprint becomes traceable through the incorporation of devices that record these movements such as sensors been placed in intersections to map car traffic, GPS located bicycles, near-field technology and cell phone trackers. These are examples of ways that data is generated that can help map the flows of people in cities.
Natural environment

Human dependency on natural elements makes it critical to monitor changes and alterations to the natural environment. Incorporating data related to this layer into the platform can include linking sensors that track and monitor information in real time related to air, water, flora and fauna. The project *Sensing city* provides an example in which several initiatives seek to give citizens the opportunity to measure different environmental aspects, such as the river quality and respiratory illnesses linked to environmental factors. The data collected is made available for academic and commercial uses (Arup 2013) (http://www.arup.com/news/2013_09_september/03_september_world_first_sensing_city_project_launched.aspx ). Natural scientists could not only access this data, but also contribute to populating the platform by uploading the data they have collected in their research.

Built infrastructure

Since the birth of cities most of the activity that goes on happens in the physical structures built such as roads, building, electricity towers and pipelines. These generate high amounts of data that could be of great help in the process to understand how cities function. Houses and offices consume electricity, gas and water and taking this data could tell us how efficient or inefficient the infrastructure is. Increasingly, sensors are integrated in the infrastructure making data collection a much easier process. This offers a good opportunity to connect the data provided by the instant communication between these data providers and the platform.

Information society

The introduction of telecommunication technologies has given way to interactions with urban environments in ways never seen before. Smartphones leave a trace of places, pictures taken, and opinions. These are important sets of data that gather the collective conscience and culture and provide valuable insights for pattern recognition.

The use and analysis of social media also give insights on human behavior and activities. For instance, the data of foursquare is an application that provides the opportunity for its users to check-in to places and is a valuable tool for market analysis of trending places. The accelerating trend in becoming an information society only facilitates the integration of data into the platform and, as a matter of fact, reinforces the need and desire to have such platform.
Economic aspects

Economic factors are important to urbanism as they are usually a key element in the growth of cities. Economic transactions are nowadays done digitally through the use of electronic cards, electronic wallets, and online transactions. These new tools allow observing how money moves in the city. Though a sensitive subject in terms of privacy, this data is important to understand the trends and preferences of citizens. This data can be incorporated into the platform with the collaboration of banks and other financial organisms managing this information.

Social aspects

The interactions of people in within the city are issues that are important to understand how the city functions. These interaction include citizens’ activities professionally but also for leisure. For example data such as crime, poverty, education, housing, and civic involvement are data sets that are recorded by city agencies and now published by the open government movement where it is available to access all this data. Also with the help of the web, people can upload data that shows relations of humans with the urban space, such as data where people find dangerous dogs or specific spots where bicycle accidents are recorded. Data related to where do people eat, or the mapping of cultural attractions is recorded by people placing their own data from social media platforms.

V. Urban informatics and the emerging science of cities

A city is considered to be a very complex system full of relations and interactions of the different layers. Lately, this complexity has been the attention of many scholars and researcher who are taking on the task of making sense of the patterns that compose it. From this, the emerging field of urban informatics was developed as the study of the varied layers that make up the city by analyzing and visualizing data that contributes to seeing the city in a more accurate manner.

The Center of Advanced Spatial Analysis (CASA) of the University College London is a center specialized in gathering data and analyzing it in ways that allow researchers to study cities in a scientific approach. As one of the top researchers Michael Batty argues, “[c]ities must now be looked at as constellations of interaction, communications, relations, flows, and networks,
rather as than locations, and argue that location is in effect, a synthesis of interactions; indeed, this concept lies at the basis of our new science.” (Batty, 2013)

In another effort, the theoretical physicist Geoffrey West and his team at the Santa Fe Institute tried to get as much urban data in all possible aspects in order to find the correlations that cities have and how they work, in a way to find the “DNA” of cities.

In the optic to contribute to the sciences of cities, the proposed platform pretends to close the gap between the data generated and its analysis. The availability and storage of the data in real time and in big quantities (varied) generates a footprint of the city. This footprint will allow researchers to further and analyze in detail all the layers that make up the urban space. This helps getting a more precise analysis of the “DNA” of the city as well as giving the capability of doing it instantly. Furthermore, the platform gives the opportunity to visualize the data in a speedy manner allows for in-depth studies and the correlation among different layers.

VI. Case studies

There have been several instances in which a number of data sets have been aggregated in order to have the information centralized to make a more efficient use of it. This section will present four examples in which this is put into practice. It should be noted that the objective of this section, or this paper for that matter, is not to come up with something completely new, but rather to assess what has been done and to build something based on past examples to understand areas of improvements and find potential synergies.

Operations Center in Rio de Janeiro Brazil

One of the initiatives is the result of a partnership between the city of Rio de Janeiro and technology company IBM, which consisted in the creation of an information management center integrating data from 30 agencies, both from the public and the private sectors. In 2010, what originally started as a response to the city’s incapacity to react to natural catastrophes such as floods and landslides in order to have a better approach to emergency response coordination, now incorporates elements to manage increased traffic and monitor weather, aggregating many layers of information and integrating them in a Google Earth View system to visualize it on an 80 square meter screen.

In addition to gathering data from different agencies relating to traffic, weather, transportation, security and others, IBM developed a program to forecast weather and floods.
Furthermore, the center collects real time streaming video provided by the 900 cameras installed throughout the city for operators, who are actually representatives of various agencies, to monitor the city’s movements. The forecasting function allows for relevant teams to be prepared ahead of time and to know what areas will need intervention. Furthermore, the real-time aspect of it makes it a much more accurate way to monitor data and find patterns to make predictions for better-adapted policy-making.

The center is directly connected to the mayor’s office and emergency response organisms and it is the first in the world to “integrate all the stages of a crisis management situation: from the prediction, mitigation and preparedness, to the immediate response to events, and finally to capture feedback from the system to be used in future incidents”. (http://www.ft.com/cms/s/0/a1f11494-1439-11e3-9289-00144feabde0.html#axzz37Ffe6Wbu)

Figure 2: Rio de Janeiro’s operations center (http://www.cidadeolimpica.com.br/wp-content/themes/cidadeolimpica_v3/projetos/EOM/en/projeto-cor.php)

Open Data for cities

Similar to the Rio de Janeiro Center of Operations, cities around the world have released their data on platforms called “Open Data”, in which public data from different city agencies and organizations is released in machine readable data sets and available for public use. This is part of an initiative by cities to improve the accessibility, transparency, and accountability of city governments.

In the city of New York, agencies routinely collect information on streets, buildings, infrastructure, businesses, and other entities within the city, including permits, licenses, crime-related data, and 311 complaints. The Mayor’s Office of Data Analytics (MODA) centralizes this data thus uniting previously disconnected pieces of information from various agencies to create a
comprehensive city-wide data platform that serves both as a record of city activity and as a foundation for New York City Open Data.

Once the data sets are unified and released as open data, anyone can access them to conduct research and analysis. Moreover, cities organize Hackathons, which are events gathering computer programmers, software developers, graphic and interface designers, and project managers. Participants collaborate to create applications and visualizations based on collected data that translate it into coherent information available to citizens.

In a Hackathon in Austin, Texas an application was developed to collect all the data of trash and recycling pick up routes and schedules. Through the geo-localization of smartphones, users have a way to know when is their trash is going to be picked up. Also, in Oakland, California a group of hackers plotted the city’s crime data showing on a heat map the places in the city where crimes were more predominant.
London dashboard

With the amount of tools and resources to get information about a city, it is easy to get lost. Just thinking about the needs of regular citizens who want to know the weather forecast, traffic status or local news shines a light on all the information desired and the time required to consult them separately on a daily basis.

A project developed by the CASA (Center for Advanced Spatial Analysis) research lab at the University College London have put in London and other UK cities an instrument that aggregates simple spatial data and displays it on a dashboard and a map. (http://citydashboard.org/london/) The CASA project is an example of the possibility of visualizing different data sets that concern the urban space.

London has the Dashboard with the most information and retrieves information from a number of sources including the Department for Environment Food and Rural Affairs, National Oceanic and Atmospheric Administration, BBC, London School of Economics, Yahoo! developer network, Port of London Authority, Transport for London, Yahoo! finance, UCL CASA, MapTube, ScotRail, Twitter. All this information is presented in a single page and a map, a useful tool used for people who are not data experts but who wish to have geo-localized data that is quick to access and easy to understand.
Figure 4: London Dashboard and map (http://citydashboard.org/london/)
WikiCity (MIT)

MIT conducted a research called WikiCity which aims to create a common format for the interchange of real-time location-based data and a distributed platform able to collect, manage, and provide such data in real-time. This is done by borrowing the concept of a “Wiki” in the sense that anyone is able to access, upload, and modify data on this system platform.

The researchers at MIT exploit the opportunities presented by a centralized platform for real-time data, including how data can be combined in response to a user's query and the reliability and trust of data sources. Furthermore, researchers emphasize the potential of data availability for the decisions of everyday activities. For instance, in the decision of going for a jog, the user can select certain parameters on the platform related to air quality, pollen in the air, and traffic. Based on this, the jogger will receive options showing the best suitable path. As such, WikiCity “is not limited to representing the city, but also instantly becomes an instrument for city inhabitants to base their actions and decisions in a better-informed manner” (Ratti 2006).

The first implementation of a WikiCity interface was attempted in Rome, Italy, in which a screen in a public plaza was projected showing a series of real-time data. The projection showed a satellite picture of Rome with overlays indicating real-time cellphone activity, real-time location of public buses, starting and ongoing event tags at their corresponding location, and live news feeds.
VII. Critical Analysis of Centralized Platform

Based on the description of the proposed centralized and the presented case studies that illustrate the implementation of different aspects, this section focuses on a critical analysis of the concept. It will first portray the raison d'être of the platform highlighting the justification for its creation while the second part depicts the limitations and challenges that arise in the construction of such platform.

Benefits

As highlighted by the case studies, data aggregation has been increasing not only due to enabling technology, but also because of the demand and the collective awareness of the opportunities provided by data aggregation. As observed, these benefits do not stand in isolation of one another and are built upon and interconnected with one another.
Cross-field analysis

One of the greatest advantages of having data sets in one location is the ability to carry out cross-field analysis by combining two or more layers of urban data. Indeed, one of the most important yet challenging aspects of urban analysis is the consideration of different variables, whose interactions have an effect on each other and thus cause a chain reaction in the whole urban system. Moreover, understanding the interactions among variables becomes a crucial step for comprehending the functioning and true nature of cities.

Real-time information

Unlike traditional data sources such as surveys that remain static and quickly become obsolete, real-time data provides a constant and updated flow of information accurately reflecting the reality. This instantaneousness gives way for increasingly adapted solutions to problems. Furthermore, assessing and evaluating implemented projects becomes much more effective because of the possibility to observe successes and challenges immediately. Consequently, planners can modify interventions accordingly and not wait until the project is completed and cannot be reversible.

Visualization

Data interpretation is always easier when presented in a graphic manner. The great amount of information hosted on the platform would make it difficult to grasp meanings of urban data. When speaking of data visualization it is important to remember that different layers of data must encompass the time and space dimensions. For urban studies, it is obvious that the spatial representation of data is integrated with cross-field, real-time data that data visualization allows for the identification of patterns that would otherwise remain invisible. Once patterns are recognized, it is possible to act accordingly rather than on irrational instincts.

Decision-making

The aggregation of data provides decision-making opportunities at different scales: from a regular citizen making his/her daily-life more efficient to an architecture firm doing a site analysis to a policy-maker implementing large scale projects. The platform enables bottom-up processes because citizens know that the data they incorporate can have an impact on the top-down decision-making. An informed decision-making process can in turn contribute to more efficient and sustainable cities.
User-friendly

In an increasingly data-driven world, a variety of stakeholders, not just coders and data experts, should be able to access data. As a platform allowing virtually anyone to enter and get data, input and extraction should be intuitive and easy to use. Having a complicated system puts at risk the very existence and success of the platform and compromises the potential uses it can provide. Having an intelligent search engine incorporated into the platform ensures that users can find the desired information quickly and efficiently, even if the user does not specify exactly what data is required.

Model creation

With a variety of data sets it becomes an advantage in the creation of models for urban simulation and predictions. It can help decision makers create models for simulation before the implementation of certain policies and projects. Models can be constantly updated with feed of real-time data, as a result being able to have instant feedback.

Limitations

The availability of all urban data has an advantage in helping people better understand the city, but also raises a series of concerns such as the possibility of the platform to become a tool for technocratic governance approach for decision-making. Decisions at a city scale should not be left only to computer based answers, the platform should be used a tool and not as dictator.

Since data is produced at an ever-expanding rate, the storage of this information becomes an important consideration. Physical data storage centers are vital for the functionality of the platform but require greater management and resources as they can be costly and require specialized technicians.

Another challenge is posed by the need to find a common language in which data (structured and unstructured) is processed and visualized, in order to create a culture of usage of the platform. Establishing this language and streamlining it in a simple and intuitive manner so that all users, real and potential, can use it is not an easy task. Yet this is a basic and necessary element in a platform in which collaboration is a foundation.

The model of the platform should be structured in an organic manner, which allows flexibility for all cities to create their unique systems and to avoid standardization of platforms.
While it may be tempting to create a one-size fits all, each city has its own set of complexities and it would be erroneous to analyze a city through the lens of another one.

The lack on anonymity is a problem since data could be traced to a single person. Technology must ensure that data input as well as the output, in its visualized form respect the privacy of individuals but at the same time provide the information necessary for urban studies.

VIII. Concluding remarks

We are in a time in which data is created in an exponential way and it is very difficult to absorb and make sense of it all. Centralizing the data makes the analysis more efficient; but the nature and origin of data means that it must be done through a collaborative process. Since there have been previous initiatives to attempt this, it is important to build upon what has already been constructed in order to continue improving data analysis. In this context, it is key to provide an interface in which people can interact with the data in an easy and intuitive manner, thus giving everyone the opportunity to be an active participant of the science of cities
**IX. Bibliography**


Supporting Urban Informatics through a Big Data Analytics Online Workbench

Authors
Christopher Pettit\textsuperscript{1} cpettit@unimelb.edu.au; Robert Stimson\textsuperscript{1} rstimson@unimelb.edu.au
Marcos Nino-Ruiz\textsuperscript{2} marcosnr@unimelb.edu.au
Luca Morandini\textsuperscript{2} luca.morandini@unimelb.edu.au
Ivo Widjaja\textsuperscript{2} ivow@unimelb.edu.au
Phil Delaney\textsuperscript{1} phillip.delaney@unimelb.edu.au
Martin Tomko\textsuperscript{2} tomkom@unimelb.edu.au
Richard Sinnott\textsuperscript{2} rsinnott@unimelb.edu.au
Bill Randolph\textsuperscript{1} b.randolph@unsw.edu.au
Tom Kvan\textsuperscript{1} tkvan@unimelb.edu.au

\textsuperscript{1} Architecture, Building and Planning Faculty - University of Melbourne, \textsuperscript{2} Faculty of Engineering,- University of Melbourne; \textsuperscript{3} City Futures - University of New South Wales

Abstract

In 2007 the world’s population became more urban than rural, and, according to the United Nations, this trend is to continue for the foreseeable future. With the increasing trend of people moving to urban localities - predominantly cities - additional pressures on services, infrastructure and housing is affecting the overall quality of life of city dwellers. City planners, policy makers and researchers more generally need access to tools and diverse and distributed data sets to help tackle these challenges.

In this paper we focus the online analytical AURIN (Australian Urban Research Infrastructure Network) workbench, which provides a data driven approach for informing such issues. AURIN has developed a portal that provides (programmatic) online access to large scale distributed and heterogeneous data resources from the definitive data provides across Australia. Many of these data providers have regional (State) data hubs to coordinate and deliver data sets to the centralized AURIN portal and it’s associated e-Research infrastructure. This includes, for example, more than 20 years of longitudinal housing data across Australia with information on each housing sales transaction at the property level. For the first time urban researchers from various universities and government organizations across Australia can now systematically access housing data to run spatial-statistical analysis to understand the driving forces behind a myriad of issues facing cities, including housing affordability which is a significant issue across many of Australia’s cities.

Keywords: e-infrastructure, spatial decision support systems, housing data
1.0 Introduction

The world’s growing and increasingly urbanized population presents significant challenges for planners and policy makers to address. In 2007 the world’s population became more urban than rural, and, according to the United Nations, this trend is to continue with the percentage of population residing in urban areas to be 59.9% and 67.2% in 2030 and 2050 respectively (United Nations 2011). The total population is projected to be 8.3 billion people and 9.3 billion people in 2030 and 2050 respectively (United Nations 2011). This growth and the increasing trend of people moving to urban localities- predominantly cities - brings about additional pressures on services, infrastructure, housing, transport and the overall quality of life of city dwellers. Issues such as housing affordability are becoming increasingly pressing for city planners and policy makers to address. This is where *urban informatics* can assist in providing data driven evidenced policy and decision-making.

Urban informatics can be defined as:

“… The collection, classification, storage, retrieval, and dissemination of recorded knowledge of, relating to, characteristic of, or constituting a city.” (Foth 2009; p23).

Over the last 50 years an increasing array of digital data has been produced relating to urban settlements resulting in the rise of the ‘information city’ as referred to by (Castells 1989). Some 80 percent of this data can be given a location attribute and can be used to create spatial databases which can then be analyzed and visualized to help understand urban growth and development and to plan for sustainable urban futures. However, as Townsend (2013) points out, we must be conscious of the limitations of our ability to predict the future and to use such information in a way to engage our communities and support bottom-up participatory planning of or cities.

In this paper we introduce a ‘big data’ analytics workbench where data representing Australian cities can be accessed, analyzed and visualized. Since 2011, the Australian Urban Research Infrastructure Network (AURIN) has made available an online (geo) workbench that provides access to over 1,000 spatial datasets from over 30 data providers from across Australia. As of June 2014 there are over 6 billion data elements that cover all major cities of Australia, crossing health, housing, transport, demographics and other essential characteristics of cities. This includes historical data, current data and future data, for example, the expected population growth for major cities.
2.0 Urban Big Data Analytical Workbenches

As discussed by Stimson and Pettit (2014) the new urban technologies that are rapidly emerging already produce massive streams of data that we refer to as ‘big data’, including data in real time and space facilitated through the proliferation and spread of wireless technology. This, along with the trend towards ‘open data’ is facilitating what Townsend (2013) refers to as the ‘technology of inclusion’. (Boima and Bonfa 2012) suggest that the ‘smart cities’ concept that is now popular and widespread is largely a result of such huge gains in computing and the associated emergence of ‘big data’. They say:

“… computer hardware and software now allow the development of huge databases and complex manipulations, which in turn increase our analytical capacity to solve complex problems. The development of cloud technology and Big Data Analytics, in particular, allows real time data (space and non) management in a multi-dimensional context to address [important and complex] issues.”

Zhen (2013) says that ‘big data’ implies what he calls ‘rich knowledge’ about a city that can help it to tackle major issues when used correctly. He refers to this as ‘urban computing’, defined as:

“… a process of acquisition, integration, and analysis of big and heterogeneous data generated by a diversity of sources in urban spaces, such as sensors, devices, vehicles, building, and human, to tackle the major issues that cities face” ... it … “connects unobtrusive and ubiquitous sensing technologies, advanced data management and analytics models, and novel visualization methods, to create win-win-win solutions that improve urban environment, human life quality, and city operation systems.”

Anderson (2013) Says ‘big data’ represents “game changing opportunities” as we enter what he refers to as the “petabyte age”, and Malvey et al. (2013, p15) refers its potential “transformational impact”. But Hilbert (2013) makes the important point that:

“… the crux of the ‘Big Data’ paradigm is actually not the increasingly large amount of data itself, but its analysis for intelligent decision-making”. (p. 4)

He suggests that ‘Big Data Analysis’ might be a more fitting term as:

“… the key feature of the paradigmatic change is that analytic treatment of data is systematically placed at the forefront of intelligent decision-making. The process can be seen as the natural next step in the evolution from the ‘Information Age’ and ‘Information Societies’ to ‘Knowledge Societies’: building on the digital infrastructure
that led to vast increases in information, the current challenge consists in converting this digital information into knowledge that informs intelligent decisions. (p. 4)

‘Smart cities’ research using ‘big data’ is also enhanced through powerful advances in Geographic Information Systems (GIS) technologies and what Gilroy (2014) refers to as:

“... the availability of consistent and accurate detailed geographic information [that] is a key enabler for the growth of national economies.”

Certainly advances in data modelling capabilities using ‘big data’ - such as data mining and large-scale simulation models and agent-based techniques - have considerable potential to enhance urban research. According to Gartner (2012) big data is now at the peak of the hype cycle and is receiving considerable attention by both business and governments (Figure 1).

Figure 1. Big Data currently at the peak of the Gartner Hype Cycle¹

---

In order to harness the deluge of big dataset and the growing number of open government datasets being made available globally, a number of companies have developed capabilities in the smart cities and related urban big data analytics arena. Probably the most known of these is the IBM Smart Cities initiative, where they have established a smart cities dashboard as illustrated in Figure 2 and predictive analytics capability working with cities such as Rio de Janeiro, Brazil, and the City of Portland, Oregon in the United States. Another dashboard that enables comparative analysis of select number of selected cities from around the world is made available through the Urban Observatory project, as illustrated in Figure 3. Other initiatives include: Cisco’s Smart+Connected Communities (http://www.cisco.com/web/strategy/smart_connected_communities.html), Siemens’ Smart City (https://www.cee.siemens.com/web/at/en/csb/CVC/Your_Industry/smart-city/Pages/smart-city.aspx) and Intel’s Sustainable Connected Cities (http://www.cities.io/).

There are also a number of significant university centers and projects being established, which are focused on the challenges of urban big data. These include, most notably the Centre for Advanced Spatial Analysis at the University College London established by Prof Batty in 1995. CASA has developed the London Dashboard, which provides a number of real-time data feeds from across the city and makes this available through an easy-to-use interface, as illustrated in Figure 4. Other significant centers include: the Urban Big Data Centre (UBDC) – University of Glasgow (http://urbanbigdatacentre.com/); The Programmable City, National University of Ireland, Maynooth (http://www.nuim.ie/progcity/); Centre for Urban Science and Progress (CUSP), New York University (http://cusp.nyu.edu); Beijing City Lab (http://longy.jimdo.com/); Future Cities Laboratory (FCL) Singapore ETH Centre, (http://www.futurecities.ethz.ch/); and the AURIN Workbench, as illustrated in Figure 5.
Figure 2. IBM’s Intelligent Operations Center for Smart Cities Dashboard.

Figure 3. Urban Observatory Dashboard for comparative city analysis: http://www.urbanobservatory.org/
Figure 4. London Dashboard – Centre for Advanced Spatial Analysis, University College London

Figure 5. Data shopping and map visualization using the AURIN Workbench
3.0 AURIN Workbench

AURIN is a large-scale national e-Research infrastructure project funded by the Australian Government. AURIN is tasked with providing seamless access to data sets and scenario modeling tools to support urban researchers and policy and decision-makers. AURIN is addressing a number of areas of societal significance, through the concept of lenses (Pettit et al. 2013b). Expert Groups have been established for 7 Lenses to date including: (i) Population and demographic futures and benchmarked social indicators, (ii) Economic activity and urban labour markets; (iii) Urban health, well-being and quality of life; (iv) Urban housing; (v) Urban transport; (vi) Energy and water supply and consumption; and (vii) Innovative urban design.

The AURIN portal (https://portal.aurin.org.au) is the flagship application of the workbench which is the focus of this particular paper. The functionality within the portal enables end users (urban researchers, policy and decision makers) to discover, analyze and visualize a plethora of datasets, which can be used to support the planning of sustainable cities. It is important to emphasize that AURIN provides programmatic access to these data sets (and the data remains as close to the data custodians as technically possible) and hence it is far more than a list of data sets that have been donated and made available through a web site (as is often the case in open data initiatives). The workbench, comprising the end user portal, is conceptually illustrated in Figure 6. It has been implemented using an open source federated technical architecture (Sinnott et al. 2012). The federated data structure enables datasets from across different cities, government agencies and the private sector to feed into the workbench. Key to the success of AURIN has been its engagement with data providers and associated stakeholders from government, industry and academia. Many organizations make available sensitive (unit level) data sets where access has to be strictly controlled and monitored. Many of these data sets have either been siloed behind organizational firewalls or have been scattered around the web. AURIN has provided the technological solution that has amassed a critical mass of data and tools that underpins a range of urban research and challenges facing our cities.
AURIN has delivered a multitude of modeling tools to support urban informatics including for example a Walkability Toolkit and a What if? Scenario Planning Tool. The Walkability toolkit includes an agent based approach for calculating pedsheds (Badland et al. 2013). The What if? Scenario Planning Tool is integrating the well known (Klosterman 1999) What if? GIS based planning support system (PSS) into the AURIN portal (Pettit et al. 2013a). Other tools within the AURIN portal include an employment clustering tool, and a suite of spatial and statistical routines, charting and mapping visualization capabilities. These tools have all been developed in an open source secure portal which is extensible and adaptable to including additional tools and dataset as they become available. The tools themselves leverage Cloud technologies through the NeCTAR project (www.nectar.org.au) and hence can scale horizontally.

All of these tools require access to a rich tapestry of datasets. The project have established a number of Data Hubs and feeds across the country to programmatically access data required to support urban researchers (Delaney and Pettit 2014). These Data Hubs close the loop between data owners and data users by ensuring each hub is established aligning to a set of core principals, defined as: (i) Facilitating collaboration and interaction between end users and data custodians (Figure 7); (ii) being held as close to the source as possible;
(iii) set up to serve a broad end user community, not a single project; and (iv) sufficient information (including metadata) being provided for users to understand the data.

Figure 7 – Data hub conceptual communication feedback cycle (Delaney and Pettit, 2014).

In this paper we introduce one of the Data Hubs which provides programmatic access to a rich set of property data available at both aggregate census geographies and for each property in Australia. The dataset is spatial-temporal and goes back over 20 years, with monthly updates and includes descriptive information such as dwelling type (house, unit, land), number of bathrooms and bedrooms and dwelling features (including more than 20 associated dwelling characteristics such as laundry room availability, garage, BBQ facilities, harbour or beach view, etc). The data also includes comprehensive information of sales, auction and rent economic cycles, including total transaction prices (with corresponding statistical analysis. For example, median price, detailed price breakdown, standard deviation price, geometric mean price, 45% quartile, etc.), as well as first and final advertisement prices and dates, settlement dates, and auction clearance rate among others. This big data is currently used for analyzing housing affordability and trends in Australia using the suite of tools available via the AURIN workbench. Figure 8 illustrates the Median House Price in Melbourne using the property data in the workbench.
This data hub aligns perfectly with the principals set out above, as the data is held directly in the Fairfax Australian Property Monitors (APM) infrastructure and queried on demand. In addition, advice and direction on the data and attributes made available was taken from a user group both within and outside of the AURIN Urban Housing Lens expert group to ensure the data made available meets existing research needs. These tools and data can be used to tackle research topics such as:

- Analyzing house price volatility: creating a house price index for various types of properties and for various sub-markets, and then using the index to model volatility in house prices and the role that the economic cycle plays in house price movement. This can be done with workbench providing data on residential properties at different time points, including spatial information about property location.

- Analyzing landscape features impact on housing prices, e.g. the differences the pricing of waterfront properties with non-water front properties during different phases of the residential property cycle, or the effects of a new train station on the surrounding house value over the time of announcement, construction and service.

- Comparison of property prices locally versus nationwide trends, e.g. conducting research on property prices in Canberra during the 1990s, and comparisons with prices nation-wide during the same time span.
• Property development for urban fringe over 20 years, e.g. supporting longitudinal research on property prices development in the urban fringe, with the provision of unit level transactions data for residential lots in the growth areas in local government areas of interest over the last 15-20 years.

• House prices change after natural disasters, e.g. research on the impact in housing prices before and after the Brisbane floods of 2010/11. Information for houses in a list of suburbs regarding whether it was damaged or not, and how much damage was sustained due to the floods.

3.1 AURIN Software Architecture and Support for Distributed Data Access

The AURIN Software Architecture follows a client-server, service-oriented architecture model applied in a fashion that maximizes re-use, scalability, and independence of individual components. The aim is to establish a loosely coupled, flexible and extensible service-based architecture (e-Infrastructure) that allows seamless, secure access to distributed data sets from the definitive urban data providers across Australia. Key to this is the consistent specification and implementation of the component APIs (Application Programming Interface). These APIs can in principle be accessed and used by external third party software. Discussions are ongoing on the more public access to these APIs.

Individual components of the AURIN e-Infrastructure communicate through Web Service API calls, typically (but not exclusively) applying the Representational State Transfer (REST)-ful style of Web services. The AURIN e-Infrastructure leverages JavaScript Object Notation (JSON - json.org) for the encoding of the majority of its communication and data. JSON allows for hybrid messages with adaptive content. This is particularly advantageous for the complex data descriptions and formats to be passed around within the AURIN e-Infrastructure. The GeoJSON extension of JSON (www.geojson.org) was adopted for internal spatial data transfers and continues to be the format of choice.

The high-level overview of the technical architecture of the underlying e-Infrastructure is shown in Figure 8.
A core mission of AURIN is to provide access to a range of federated data sources from an extensive and extensible range of data providers. The e-Infrastructure has therefore been designed in a flexible manner to support access to and ingestion of diverse data sources in a unifying environment. It was a key requirement to offer access to the e-Infrastructure (data and tools) with little or no demands on the client (user) environment. The sole demand placed on clients/users is a reasonably modern browser. The recommended browsers include Firefox, Chrome and Safari. (Internet Explorer is deprecated due to its non-standard handling of certificates).

Access to the e-infrastructure is delivered through the Australian Access Federation (AAF). AURIN offers single sign-on to all of the data sets, services and tools offered by all providers. At present, the portal has been (successfully) used by over 30 Identity Providers (users) across the AAF, i.e. one or more users have logged in from those sites. This includes all major research-oriented organizations across Australia (universities and research organizations such as CSIRO). It is important to note that single sign-on here does not imply that all data and tools are accessible to all users at all times. Several data providers place specific demands on who can...
access their data and tools and what they can subsequently do with this data. Fairfax Australian Property Monitors (APM) and the Public Sector Mapping Agency (PSMA) are two examples of industrial partners that require such restricted usage of their extensive data holdings with policies on download established and enforced. In this case, AAF (academic collaborators) are allowed access to these resources whilst non-academic collaborators are restricted, i.e. those who authenticate through the AAF Virtual Home Organisation (VHO).

To provide this finer grained access control, the e-Infrastructure has established an Edge Security service. This component of the architecture extracts information from the Security Assertion Markup Language (SAML) attributes delivered by the AAF through the Internet2 Shibboleth protocols to the portal, and subsequently uses this to define the access and use permissions for that individual on services, data and tools that are accessible. Thus VHO authenticated users will not see restricted data sets. Refinements to the Edge service for further, finer grained security needs are ongoing.

The typical lifecycle of data handling through the e-Infrastructure requires several core capabilities. The registration of a dataset (federated or local) in the system utilizes the Data Registration Service which includes support for enriched metadata capture necessary for the advanced handling of the data, e.g. by the user interface and workflow/analytical capabilities. A range of refinements to the Data Registration Service have been undertaken with specific focus on its overall usability for end user data providers (who are required to provide the necessary metadata).

The AURIN Data Provider Service focuses on the interactions with remote data providers to access and deliver their data sets to the e-Infrastructure. The Data Provider Service supports a multitude of protocols, interface flavours and must deal with a diverse range of data formats. These include geospatial data providers (typically using Open Geospatial Consortium Web Feature Services (WFS), ReST based Services, and in the case of the ABS Statistical Data Markup Exchange (SDMX) data services). Conversion of data from remote data providers to internal data formats (JSON+GeoJSON) that is used across the internal components is a key part of the functionality offered by the Data Provider Service. The Data Provider Service now includes a range of new clients and has been systematically hardened to support larger data sets and data sets of novel structures, for example data cubes to handle multi-dimensional data such as the APM housing data and other significant data holding including those from the Australian Bureau of Statistics (ABS) and the National Socio-Economic Modelling (NATSEM) Centre, University of Canberra. (Widjaja et al. 2014)

The AURIN Persistent Data Store is used when requested (shopped) data needs to be stored in a persistent manner, e.g. for further analysis. The data store has been realised with the noSQL document-based solution.
CouchDB. This system includes support for larger data sets and data joins. The data itself is still stored as JSON+GeoJSON objects. The noSQL structure of the database allows for flexible storage of datasets that adapt to the requirements of the metadata that need to be stored for each particular type of data source. The AURIN system has been specifically designed to cope with this schema-less approach. This allows tackling the extremely heterogeneous data sets being delivered by distributed data providers. The persistent AURIN Data Store itself offers a ReST-based API supporting storage of JSON+GeoJSON formatted objects and subsequent user access to these datasets.

In addition to CouchDB, a relational database (PostGIS) is used within the e-Infrastructure for dealing with high performance storage and querying of structured datasets, in particular those with spatial-geometric attributes. This is required especially for projects such as What-If? which require relational data management solutions. PostGIS provides an extension to Postgres for spatial data handling. Topologically accurate data sets are incorporated into back end regionalisation databases. These are generalised to the user interface (removing some of the accuracy of information as displayed in the browser to avoid large scale geospatial boundary data transfer to/from the client browser thereby impacting directly on the user experience).

Other e-infrastructures components comprising the technical architecture include the middleware, APIs, workflow, user interface and software production environment. These are beyond the scope of this paper to discuss in further detail.

3.2 Housing Data Clients

AURIN supports an extensible range of targeted data clients for accessing and delivering data according to the remote data provider demands and the format and nature of the data. The data clients are implemented using open source standards and technologies common to all components in AURIN. The housing data hub client components are based on a Java-based platform instrumented through the Spring API2. This client implements a Representational State Transfer (ReST) web services that is used to interact with both the modular infrastructure of AURIN and the data infrastructure of the data provider – the Fairfax Australian Property Monitors (APM) – one of the main providers for housing information across Australia.

The client supports the following capabilities and services:

2 http://spring.io/
• Discover web service: displays the relevant information that relates to the type of meta data that is being requested.
• Property Level Event Data web service: delivers the individual property level event data for one user defined date range, including full address and cadastral identifications.
• Aggregated Statistic Data web service: provides access to the APM aggregated statistic data. APM updates the aggregated data sets on a regular basis to utilise the newly collected data for the statistic data of the recent property market (availability from two weeks until 20 years in the past for some data sets).
• Disclaimer web service: delivers the APM Terms and Conditions web service to support contractual and legal compliance, which every user must accept to use this information.

Each type of the aggregated statistic data is defined by geographical, chronological and property categories. The web service allows drilling down to each pre-agreed combination of each of the above-mentioned categories. The web service also supports further filtering of the result set by a geographic bounding box or polygon, and/or by a time period.

The data itself relies on APM’s existing data importing pipeline for the property level event data, and the extension to the APM existing statistic calculation process for monthly compiling and updating the row level and aggregated statistics data.

It supports five statistic types:
  o General Sold
  o For Sale
  o For Rent
  o Auction
  o Private Treaty Cycle

For the aggregated data service, each type of statistics has three dimensions:
  o Geographical Categorisation (Suburbs, Census Districts, Local Government Areas, etc.)
  o Chronological Categorisation (one months span, three months span, yearly, etc.)
  o Property Categorisation (House, Unit, Land)

For the row level (property) data service, each type of statistics has three dimensions:
o Geographical detail (Full address disclosure, Meshblock)
o Chronological timeframe (complete sales records for any defined time period from 1993 until 2 weeks prior to the present date)
o Property Categorisation (House, Unit, Land)

3.3 Data Hub Interfaces

Extensive collections of federated datasets, visualisations and analytical tools are available through the AURIN portal. The portal makes use of select-retrieve-visualise-and-analyse user interaction paradigm. In this paradigm, users first need to select a study area from various regionalisations within AURIN Geo-classification services. Users then can search and pick a dataset from our Data Registry. Users can query the data catalogue by organisation, keywords, or the aggregation level of datasets. The search is also context-sensitive, means it provides datasets only available within the selected area study. User can select the attributes of interest in the dataset and specify the required parameters such as dimensional values and temporal extent of the dataset.

After the dataset of interest is identified, the user can add this into his/her project and generate a request to the data provider to retrieve the data for the selected study area currently, from any urban locality across Australia.

Once the data is retrieved and available as part of the working project, the user can visualise this on the map or as a chart. The portal supports standard map thematic visualisation (choropleth) using various categorisation strategies (natural break, equal interval, quantile), non-spatial and spatial graph visualisation, and GeoJson plot. It provides visualisations using typical charts like bar chart, line chart and scatter plot. More advanced data analytics could be carried out using various tools available with the AURIN workflow environment. The portal supports tabular data manipulation, dataset join, standard statistical analysis, spatial statistics and other specialist tools like Walkability, Migration Analysis, and Bipartite Graph Analysis (BPNET). These workflows may produce some new visualisations and new datasets.

All the datasets and the data processing results can be downloaded as CSV files, SHP files, or GeoJSON files as not all end users will want to utilise the tools within the AURIN workbench.
4.0 Housing Analytics using AURIN Workbench

AURIN has established Expert Groups comprising domain experts from across academia, government and industry for each of 10 ‘aspirational’ lenses, (with 7 of these lenses being implemented). These Expert Groups assist in determining the scope, focus and situational contact for the work of the Lenses with a focus on identifying key datasets and tools to be incorporated into the AURIN workbench (Stimson et al. 2011). Over a series of workshops, the Expert Groups have defined the framework for guiding AURIN infrastructure development. The framework developed for the Urban Housing Lens is illustrated in Figure 9. The next section of the paper will provide some examples of how the data can be used in more sophisticated ways for understanding housing affordability issues in Australia.

Figure 9. AURIN Urban Housing Lens Operational Framework (Pettit et al. 2013b).
4.1 AURIN Data for understanding housing affordability – a case study

Housing affordability is a persistent policy issue. It has also been the focus of interest to housing researchers for many years. The recent escalation in property prices, especially in Australia’s major cities, is an all too familiar feature of popular media reports and conversations around the garden barbeque. Yet it remains a policy Cinderella, with much wringing of hands accompanied by little active policy deliberation. The complexity of the drivers of housing markets and the range of approaches to explaining these processes militate against a coherent understanding these drivers or proscribing policy solutions to improve affordability outcomes.

Defining what is or is not an affordable home is also problematic and subject to longstanding debates. The two basic components of any housing affordability assessment, namely housing costs and household income and wealth, defy easy measurement, especially on an ongoing basis. Debate over the appropriate measurement of housing affordability is also contested. The prevalent use of ratio measures (housing cost as a percentage of household income) stems from the more readily availability of such data and the ease of interpretation. However, other methods, such as assessing the household income remaining after housing costs are deducted and comparing this to an income benchmark of some kind, might be more appropriate, especially for lower income households for whom housing affordability matters the most.

The role of data in this is critical. Understanding how and why housing markets function and whether or not housing is affordable to various sections of the population is a fundamental prerequisite for coherent policy development. AURIN’s urban housing lens is focused on delivering the kinds of data to support both systematic monitoring of housing affordability but also providing the base data for developing much better understanding of how the market works and where pressures of housing costs, and therefore housing affordability, occur and for whom.

Ironically, it is not that we do not have data to help us. In fact, in Australia, details of every residential property sale and rental agreement are gathered by State and Territory governments as on-going statutory administrative requirements. These are all recorded at address level and therefore can be geo-coded to allow a precise spatial matching to the land use property cadaster. The potential for detailed spatial analysis of these data is therefore significant. In the case of property sales, these data are sold to commercial companies for on-sale, once suitably processed, to the real estate industry, media, insurance industry, banking sector and others to assist in their business activities. There are several major national private firms that gather and disseminate sales data from around Australia on a for-profit basis. However, access to property sales data by the research and policy
community is much less ubiquitous and often incurs significant expense or time consuming negotiation, often stymied by data protection concerns. This is a paradox – while your corner estate agent can tell you the sales history of your house in some detail, a university researcher has significant difficulty in obtaining the same data for research purposes, unless he or she has the cash to buy it. AURIN’s relationship with Australian Property Monitors, who have made their national case level sales database available, is a major step forward in broadening access to these data.

For rental data, the problem is more intractable as it is held by the various jurisdictional Rental Bond Boards who have no direct interest in the market information the records contain and, to date, have shown little interest in disseminating it. And unlike house price data, as yet no-one has attempted to assemble a nationally consistent dataset on rents from the various jurisdictional departments that gather this information. Other data, from Federal rent support and first home owner payments to mortgage and foreclosure data, are all collected by various government and private agencies at address level. Assembling these data together in a nationally consistent manner, geocoded to match with the property cadaster, would provide researchers and policy makers with a vastly greater capacity to study and better understand how housing markets function.

Until now, the barriers to the assemble and dissemination of these kinds of housing data means that the measurement and analysis of housing affordability in both home ownership and rental sectors remains poorly developed, especially at a disaggregated spatial scale. The latter is a critical gap, given the spatial complexity of housing sub-market. There are national affordability measures, largely issued by major banks, which purport to chart average home ownership affordability trends. But these are comprised of aggregated statistics based on simplifying assumptions that lack spatial sophistication. And nobody monitors rental affordability trends in a comparable manner.

AURIN has the potential to change all this. The first phase of AURIN’s housing lens activity has been to develop a number of pilot projects that establish the viability of assembling address level and other housing data from a variety of sources and how these might be integrated with synthetic measures of housing affordability as well as new measures developed using the address level sales and rental data being brought together for the first time. These will be augmented by small area Census data on household incomes disaggregated by household characteristics and dwelling profile. The much greater sub-market analysis of housing affordability across the country that this integration promises to open up will propel our capacity to monitor and study housing affordability at both the macro and local scales across our urban areas. Figure 10, using sales data from the NSW Valuer General’s data base and ABS Census, which are available through the workbench, illustrates how address level sales data can be used to develop a measure of housing affordability
that allows the mapping and analysis of homes that are sold that are affordable to moderate income families in three time periods. These data can be analysed and presented at a range of spatial scales and time periods, allowing an ongoing monitoring of housing affordability outcomes.

Once this basic capacity is established, additional analytical tools, such as spatially sophisticated hedonic pricing models, will become more easily developed. New spatially disaggregated modelling of demographic and income driven urban growth models could be developed with implications for the location of new dwelling construction. New analytic tools for visualizing the data, including flow data relating to household movements in housing sub-markets are also being piloted. A range of basic housing market indicators, such as rental yields, vacancy rates, turnover rates and other indicators of housing market dynamics will be possible from the address level data. In time more sophisticated modelling of the nexus of sub-area housing demand and supply will become possible.

2001

2006
5.0 Conclusions

Affordable housing is but one part of a larger societal goal of affordable living (Kvan and Karakiewicz 2012). The parameters for the latter are necessarily more diverse that those that contribute to an analysis of affordable housing alone. By applying factors such transportation, employment, education, recreation, access to food and healthy living inform the base model. Many of these are within the AURIN data sets; the potential of the workbench is to deliver access to the rich variety of data sets such that the larger questions of affordable living can be examined with a robust model of affordable housing underpinning the analysis.

In this paper we have discussed the workbench and the Australian Property Monitor Data Hub. Access to the workbench currently enables both urban researchers and government policy decision makers to access, analyse and visualization big urban datasets such as the property sales databases for dealing with pressing issues facing our cities such as housing affordability and ultimately affordable living. We have presented some of the core technical infrastructure components of the AURIN workbench and introduced the data hubs concept, which underpins the federated data retrieval mechanism for sourcing. Currently there are more than 1,000 datasets available via the workbench. These comprise more than 6 billion data elements available that cover all major cities of Australia. Future work in the project will see a growing number of data sets (big and little) being made available through the workbench through direct programmatic access.
6.0 Acknowledgements

AURIN is a $24 million project developing national networked urban research infrastructure capability in Australia funded by the Commonwealth Government under the Education Investment Fund and the National Collaborative Research Infrastructure Strategy. The University of Melbourne is the lead agent for the project. The authors thank their colleagues in the AURIN Office and the Core Technical Team, along with the many contributors to the AURIN project in a number of institutions across Australia, including many government and other agencies providing access to data, through Data Access Agreements and Sub-contracted projects.

7.0 References

Hilbert M. 2013. Big data for development: From information-to knowledge societies. Available at SSRN 2205145.


A Big Data Mashing Tool for Measuring Transit System Performance

By:
Gregory D. Erhardt
greg.erhardt.13@ucl.ac.uk

Oliver Lock
oliver.lock.13@ucl.ac.uk

Dr. Elsa Arcaute
e.arcaute@ucl.ac.uk

Professor Michael Batty
m.batty@ucl.ac.uk

Centre for Advanced Spatial Analysis (CASA)
University College London
90 Tottenham Court Road, London W1T 4TJ
United Kingdom
+44 20 3108 3876

Submitted to:
The Big Data and Urban Informatics Workshop 2014
11-12 August 2014
Chicago, Illinois, USA

Submitted on:
15 July 2014
Abstract

This research aims to develop software tools to support the fusion and analysis of large, passively collected data sources for the purpose of measuring and monitoring transit system performance. This study uses San Francisco as a case study, taking advantage of the automated vehicle location (AVL) and automated passenger count (APC) data available on the city transit system. Because the AVL-APC data are only available on a sample of busses, a method is developed to expand the data to be representative of the transit system as a whole, and to impute missing values. In the expansion process, the General Transit Feed Specification (GTFS) data are used as a measure of the full set of scheduled transit service.

The data mashing tool reports and tracks transit system performance in these key dimensions:

- **Service Provided**: vehicle trips, service miles;
- **Ridership**: boardings, passenger miles; passenger hours, wheelchairs served, bicycles served;
- **Level-of-service**: speed, dwell time, headway, fare, waiting time;
- **Reliability**: on-time performance, average delay; and
- **Crowding**: volume-capacity ratio, vehicles over capacity, passenger hours over capacity.

An important characteristic of this study is that it provides a tool for analyzing the trends over a significant time periods—from 2008 through the present. The tool allows data for any two time periods to be queried and compared at the analyst’s request, and puts the focus specifically on the changes that occur in the system, and not just observing current conditions.

Key Words: Transit Performance Monitoring, AVL-APC, GTFS

Word Count: 6,509 words + 6 tables and figures
1. Introduction

Performance-based planning builds upon the traditional transportation planning process by aligning planning goals and objectives with specific performance measures against which projects can be evaluated. The emergence of performance-based planning received a boost from the current U.S. federal transportation legislation (MAP-21, 2012) which makes it more central to the overall planning process. In recent years, researchers and practitioners have made significant progress in developing approaches to performance-based planning (Turnbull, 2013), including approaches to establishing performance-based planning programs (Lomax, Blankenhorn, & Watanabe, 2013; Price, Miller, Fulginiti, & Terabe, 2013), methods for converting data into performance measures (Benson, Perrin, & Pickrell, 2013; Winick, Bachman, Sekimoto, & Hu, 2013; Zmud, Brush, & Choudhury, 2013), and experience formulating relevant performance measures from institutional priorities (Lomax et al., 2013; Pack, 2013; Price et al., 2013). In spite of this momentum, a number of challenges still remain, including the availability of supporting data, the ability to synthesize those data into meaningful metrics and the resources required for analysis (Grant, D’Ignazio, Bond, & McKeeman, 2013).

This research aims to meet these challenges by developing software tools to support the fusion and analysis of large, passively collected data sources for the purpose of measuring and monitoring transportation system performance. Because they are continuously collected, Big Data sources provide a unique opportunity to measure the changes that occur in the transportation system. This feature overcomes a major limitation of traditional travel data collection efforts, which are cross-sectional in nature, and allows for a more direct analysis of the changes that occur before-and-after a new transport project opens.

This first phase of work focuses on transit system performance, with future work planned to integrate highway measures. This study uses San Francisco as a case study, taking advantage of the automated vehicle location (AVL) and automated passenger count (APC) data available on the city transit system.

As of the year 2000, automated data collection systems were becoming more common at transit agencies, but data systems were immature, network and geographic analysis methods were in their infancy, and the data were often used for little beyond federal reporting requirements (Furth, 2000). Subsequently, TCRP Report 88 provided guidelines for developing...
transit performance measurement systems, with a focus on identifying appropriate performance measures to correspond to agency goals (Kittelson & Associates et al., 2003). By 2006, TCRP 113 identified a wider range of AVL-APC applications, but still a dichotomy between APC data which was used in its archived form and AVL data which was often designed for real-time analysis and not archived or analyzed retrospectively (Peter G. Furth, Hemily, Muller, & Strathman, 2006). More complete data systems have since been developed that encapsulate the data processing and reporting (Liao & Liu, 2010; Liao, 2011), apply data mining methods in an effort to improve operational performance (Cevallos & Wang, 2008), and examine bus bunching (Byon et al., 2011; Feng & Figliozzi, 2011). Initial attempts have been made to visualize the data at a network level (Berkow, El-Geneidy, Bertini, & Crout, 2009; Mesbah, Currie, Lennon, & Northcott, 2012).

Two important characteristics distinguish this study from previous work.

First, it operates on a sample of AVL-APC data, and a methodology is established to expand the data to the schedule as a whole, impute missing values, and weight the data to represent total ridership. This is in contrast to the examples given above which generally assume full data coverage. Establishing expansion and weighting methods is important because it allows Big Data analysis to be applied in a wider range of locations with lower expenditure on data collection equipment.

Second, this study develops a tool to analyze the trends over a significant time periods—from 2008 through the present—as opposed to many applications which focus on using the data to understand a snapshot of current operations (Liao & Liu, 2010; Feng & Figliozzi, 2011; Wang, Li, Liu, He, & Wang, 2013; Chen & Chen, 2009). The tool allows data for any two time periods to be queried and compared at the analyst’s request, and puts the focus specifically on the changes that occur in the system, and not just observing current conditions. For example, changes that occur in a specific portion of the city may be traceable to housing developments or roadway projects at that location, trends that may go unnoticed given only aggregate measures or cross-sectional totals.

The remainder of this paper is structured as follows: Section 2 describes the data sources used in this study. Section 3 covers the methodology for data processing, including the approach used to expand and weight the data to be representative of the system as a whole. Section 4
presents example outputs to demonstrate the types of performance reports that the data mashing tool can produce. Section 5 is conclusions and expected future work.

2. Data Sources

This research uses two primary data sources provided by the San Francisco Municipal Transportation Agency (SFMTA): automated vehicle location/automated passenger count (AVL-APC) data, and archived General Transit Feed Specification (GTFS) data. A third data set, from the Clipper transit smartcard system, will be incorporated if access can be negotiated in a manner that appropriately addresses privacy concerns.

The AVL-APC data is formatted with one record each time a transit vehicle makes a stop. At each stop, the following information is recorded:

- Vehicle location;
- Arrival time;
- Departure time;
- Time with door open;
- Time required to pullout after the door closes;
- Maximum speed since last stop;
- Distance from last stop;
- Passengers boarding;
- Passengers alighting;
- Rear door boardings;
- Wheelchair movements; and
- Bicycle rack usage.

In addition, identifiers are included to track the route, direction, trip, stop, sequence of stops, and vehicle number. The vehicle locations reflect some noise, both due GPS measurement error and due to variation in the exact location at which the vehicle stops. However, because the stop is identified, those locations can be mapped to the physical stop location, providing consistency across trips. The count data become less reliable as the vehicle becomes more crowded, but the
data are biased in a systematic way, and SFMTA makes an adjustment in the data set to compensate for this bias. The data are not currently available on rail or cable car, only on the busses. Equipment is installed on about 25% of the bus fleet, and those busses are allocated randomly to routes and drivers each day at the depot. These data are available from 2008 to the present.

Because the AVL-APC data are available for only a sample of bus trips, the GTFS data are used to measure the scheduled universe of bus trips. GTFS is a data specification that allows transit agencies to publish their schedule information in a standard format. It was initially used to feed the Google Maps transit routing, and is now used by a wide range of applications. The data are in a hierarchical format and provide the scheduled time at which each vehicle is to make each stop. The full specification is available from (“General Transit Feed Specification Reference - Transit — Google Developers,” n.d.). The data used in this study were obtained from the GTFS archive (“GTFS Data Exchange - San Francisco Municipal Transportation Agency,” n.d.), from 2009 to present.

In addition, negotiations are underway seeking access to data from the Bay Area’s transit smartcard system, Clipper Card. These data provide the time and location of any tap-ins to pay a fare or transfer to a transit vehicle. Clipper Card was introduced in 2010, and currently has a penetration rate of approximately 30% of riders. The data provide value over the above sources because they allow transfers to be identified, and can be used to impute the boarding and alighting locations of discrete transit trips. The data are subject to California’s laws governing personally identifiable information (Harris, 2014), making data privacy and protection issues of particular importance.

3. Methodology

This section describes the methodology used to generate transit performance reports from the raw data. To ensure the performance measures are a valid representation of the transit system, the data area cleaned, expanded and weighted as outlined in FIGURE 1. The transit smartcard data is included in the figure to demonstrate how it fits with the process, although its incorporation is left to future work.
First, each individual data set is cleaned and converted into a common format. For the AVL-APC data, this involves filtering out non-revenue service, records without a valid route ID, stop ID or trip ID, duplicate records, and those that do not meet quality control requirements. A number of derived fields are added, including the arriving and departing passenger load, the schedule deviation, flags for on-time arrival, time period groupings, and end-of-line flags. All date and time fields are converted from string format to a native Datetime format that allows for easy sorting and calculation of differences. As part of this Datetime conversion, special care is taken to handle the wrap-around effects of trips occurring between midnight and 3 am, which continue the schedule of the day prior, and whose ridership is counted with the day prior. An equivalency file is read to attach route IDs consistent with the GTFS data so the two files can later be joined. As part of this cleaning and processing the data are converted from their raw text file format to an HDF Datastore format, as described later in this section.
The raw GTFS data are read and converted to a record-based format such that they are directly comparable to the AVL-APC data. This format has one record for each stop made by each vehicle trip on each transit route. These data are written separate for each day, making the identification of weekday, Saturday or Sunday/holiday service explicit. The process makes time periods, trip IDs, direction IDs and route IDs consistent with the equivalency used for the AVL-APC data. It calculates the headway of each trip, the scheduled runtime from the previous stop, and the distance traveled from the last stop, and along the route shape as a whole.

After the initial cleaning and conversion, the data are joined to create an expanded data store. The goal of this expansion is to identify exactly what is missing from the sampled data, so they can be factored up to be representative of the universe as a whole. The relationship between the data sets is that transit smartcard data provides a sample of about 30% of riders, the AVL-APC data provides 100% of riders on a sample of about 25% of vehicle trips, and the GTFS data identifies 100% of vehicle trips. Therefore, the expansion chain allows the more information-rich data sets to be combined with the more complete, but less rich data, much like a household travel survey would be expanded to match Census control totals. In this case, the expansion is a left join of the AVL-APC data records onto the GTFS records. The resulting Datastore has the full enumeration of service, but ridership and actual time information attached to only a portion of records. Without this step, it would not be possible to differentiate between trips that are missing because of a service change or those that are missing because they were simply not sampled. In a setting where we are explicitly interested in examining service changes, this distinction is important.

The next step is to aggregate the daily data to create monthly averages for weekdays, Saturdays and Sundays/holidays. The monthly resolution corresponds to transit industry standards for reporting ridership. Using weekdays as an example, the averaging is done by first selecting all days in the month where a weekday schedule was in operation (according to GTFS). Usually, there will be 21 or 22 weekdays in a month, and the non-missing observations of each of those days, usually about 5, will be averaged. This averaging is specific to the route, the direction, the trip and the stop. For example, consider the 1-California route, inbound direction, trip departing the first stop at 7:22 am, and third stop of the trip. In this way, each trip is always averaged with the same departure from other days, resulting in a Datastore that represents
average conditions, but still retains the full detail of every route, direction, trip and stop scheduled in the system. This averaging process becomes a bit more complicated in cases where the schedule changes mid-month. In those cases, it is necessary to track the number of days in which each schedule is in operation, such that the trips from each schedule can be weighted and combined accordingly.

Over the course of the month, about 90% of scheduled weekday bus trips are observed at least once, and about 65% of scheduled Saturday and Sunday/holiday bus trips are observed at least once. Rail does not have AVL-APC equipment installed, so rail remains unobserved in this structure. Summing to system totals at this point would still leave bus ridership understated by about 10% on weekdays, and more on weekends, so an additional imputing and weighting process is introduced. First, if a trip is unobserved in the month of interest, the software attempts to impute its value from using the same trip from another month. This is done by looking at the month before, then the month after, then two months before, then two months after. If the trip is not observed at all over that five month period, then the imputation process stops, and it remains unobserved. At the end of this step, about 98% of weekday trips and 93% of weekend trips are either observed or imputed. Note that this process is not able to account for scheduled trips that are not run, due to driver or equipment availability or other operational issues.

To compensate for the remaining missing trips, the software calculates a weighting factor and applies it to scale up the ridership on observed or imputed trips. To calculate the weights, the records for a specific month and day-of-week are grouped by route, direction, stop and time-of-day. Within each group, the weight is the ratio of total trips to observed/imputed trips. Making the weights time-of-day specific ensures that if a peak trip is missing, other peak trips will be scaled up to compensate, rather than scaling up off-peak trips that would likely have different ridership. Future work will measure the difference between the base data and the imputed and weighted data, and evaluate the validity of the process.

The output of this process is a Datastore whose structure is shown in TABLE 1. The table also shows the source of each field, and whether it is included in the imputation and weighting process. Records are defined by a unique combination of values in those fields identified with as an index in the source column. This weighted and imputed Datastore is then
used to create a series of performance reports, which are discussed in the next section of this paper.

The software was developed in an open-source framework in the Python environment. It is available under the GNU General Public License Version 3 for distribution (Erhardt, 2014). It leverages several open-source packages specifically designed to provide high-performance data storage, access and analysis for extremely large data sets. Specifically:

- **Pandas** is used for in-memory data operations, providing data structures and analysis tools for fast joins, aggregations, and tabulations of the data. Its functionality is similar to what is available in an R dataframe.

- **HDF5** (Hierarchical Data Format 5) is used to store the data on disk. It is designed for the fast and flexible storage of large data sets, allows for any combination of key-value pairs to be written, and allows on-disk indexing of the data.

- **PyTables** is a package for managing hierarchical datasets designed to easily cope with extremely large data sets. PyTables serves as the interface between Pandas operations in memory and the HDF5 storage on disk.

The advantage to using this combination of technology is that it allows datasets too large to be stored in memory to be written to disk, but allows for random access to those data with very fast queries. The development has shown that the converted data are dramatically faster to access than in their raw text format. This workflow also provides much greater flexibility than using a traditional database, which typically perform best with a stable data structure, making them less ideal for exploratory analysis.
### TABLE 1 Data Dictionary for Weighted & Imputed Datastore

<table>
<thead>
<tr>
<th>Category</th>
<th>Field</th>
<th>Description</th>
<th>Type</th>
<th>Source</th>
<th>Imputed</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monthly Aggregation</td>
<td>MONTH</td>
<td>Month and year</td>
<td>Datetime</td>
<td>Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DOW</td>
<td>Day of week (1=Weekday, 2=Saturday, 3=Sunday/Holiday)</td>
<td>Integer</td>
<td>Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TOD</td>
<td>Time of day</td>
<td>String</td>
<td>Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observation Statistics</td>
<td>NUMDAYS</td>
<td>Number of days in month with this DOW</td>
<td>Integer</td>
<td>Calculated</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TOTTRIPS</td>
<td>Total vehicle trips</td>
<td>Integer</td>
<td>Calculated</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>OBSTRIPS</td>
<td>Observed vehicle trips</td>
<td>Float</td>
<td>Calculated</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IMPTRIPS</td>
<td>Imputed vehicle trips</td>
<td>Float</td>
<td>Calculated</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>IMPUTED</td>
<td>Source of imputed data (-1=month before, +1=month after)</td>
<td>Integer</td>
<td>Calculated</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>WEIGHT</td>
<td>Expansion weight</td>
<td>Float</td>
<td>Calculated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Index Fields</td>
<td>AGENCY_ID</td>
<td>Agency ID (i.e. SFMTA)</td>
<td>String</td>
<td>Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ROUTE_SHORT_NAME</td>
<td>Route short name (i.e. 38)</td>
<td>String</td>
<td>Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ROUTE_LONG_NAME</td>
<td>Route long name (i.e. GEARY)</td>
<td>String</td>
<td>Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DIR</td>
<td>Direction (0=outbound, 1=inbound)</td>
<td>Integer</td>
<td>Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRIP</td>
<td>Trip ID, as HHMM of first departure</td>
<td>Integer</td>
<td>Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SEQ</td>
<td>Stop sequence within route</td>
<td>Integer</td>
<td>Index</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ROUTE_TYPE</td>
<td>Type of route (0=tram, 3=bus, 5=cable car)</td>
<td>Integer</td>
<td>GTFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Route Attributes</td>
<td>TRIP_HEADSIGN</td>
<td>Headsign on bus indicating destination (i.e. Ocean Beach)</td>
<td>String</td>
<td>GTFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>HEADWAY</td>
<td>Headway (min)</td>
<td>Float</td>
<td>Calculated</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>FARE</td>
<td>Full fare ($)</td>
<td>Float</td>
<td>GTFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PATTCODE</td>
<td>Pattern code (i.e. 38OB3)</td>
<td>String</td>
<td>AVL-APC</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCHOOL</td>
<td>Flag indicating whether trip specifically serves school children</td>
<td>Integer</td>
<td>AVL-APC</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Stop Attributes</td>
<td>STOPNAME</td>
<td>Name of stop (i.e. Geary Blvd &amp; Divisadero St)</td>
<td>String</td>
<td>GTFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STOPNAME_AVL</td>
<td>Name of stop in AVL-APC data (i.e. GEARY BLVD&amp;DIVISADERO ST)</td>
<td>String</td>
<td>AVL-APC</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STOP_LAT</td>
<td>Latitude of stop location</td>
<td>Float</td>
<td>GTFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>STOP_LON</td>
<td>Longitude of stop location</td>
<td>Float</td>
<td>GTFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>EOL</td>
<td>End of line flag (1=end of line, 0=not)</td>
<td>Integer</td>
<td>GTFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SOL</td>
<td>Start of line flag (1=start of line, 0=not)</td>
<td>Integer</td>
<td>GTFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TIMEPOINT</td>
<td>Timepoint flag (1=stop is a timepoint in schedule, 0=not)</td>
<td>Integer</td>
<td>AVL-APC</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>
TABLE 1 Data Dictionary for Weighted & Imputed Datastore, continued

<table>
<thead>
<tr>
<th>Category</th>
<th>Field</th>
<th>Description</th>
<th>Type</th>
<th>Source</th>
<th>Imputed</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Times</td>
<td>ARRIVAL_TIME_S</td>
<td>Scheduled arrival time</td>
<td>Datetime</td>
<td>GTFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ARRIVAL_TIME</td>
<td>Actual arrival time</td>
<td>Datetime</td>
<td>AVL-AKC</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>ARRIVAL_TIME_DEV</td>
<td>Deviation from arrival schedule (min)</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DEPARTURE_TIME_S</td>
<td>Scheduled departure time</td>
<td>Datetime</td>
<td>GTFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>DEPARTURE_TIME</td>
<td>Actual departure time</td>
<td>Datetime</td>
<td>AVL-AKC</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>DEPARTURE_TIME_DEV</td>
<td>Deviation from departure schedule (min)</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DWELL_S</td>
<td>Scheduled dwell time (min)</td>
<td>Float</td>
<td>AVL-AKC</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>DWELL</td>
<td>Actual dwell time (min)</td>
<td>Float</td>
<td>AVL-AKC</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>RUNTIME_S</td>
<td>Scheduled running time (min), excludes dwell time</td>
<td>Float</td>
<td>GTFS</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>RUNTIME</td>
<td>Actual running time (min), excludes dwell time</td>
<td>Float</td>
<td>AVL-AKC</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>SERVMILES</td>
<td>Service miles</td>
<td>Float</td>
<td>GTFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SERVMILES_AVL</td>
<td>Service miles from AVL-AKC data</td>
<td>Float</td>
<td>AVL-AKC</td>
<td></td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>RUNSPEED_S</td>
<td>Scheduled running speed (mph), excludes dwell time</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>RUNSPEED</td>
<td>Actual running speed (mph), excludes dwell time</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ONTIME2</td>
<td>Vehicle arrives at stop within 2 minutes of schedule (1=yes, 0=no)</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ONTIME10</td>
<td>Vehicle arrives at stop within 10 minutes of schedule (1=yes, 0=no)</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Ridership</td>
<td>ON</td>
<td>Boardings</td>
<td>Float</td>
<td>AVL-AKC</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>OFF</td>
<td>Alightings</td>
<td>Float</td>
<td>AVL-AKC</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>LOAD_ARR</td>
<td>Passenger load upon arrival</td>
<td>Float</td>
<td>AVL-AKC</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>LOAD_DEP</td>
<td>Passenger load upon departure</td>
<td>Float</td>
<td>AVL-AKC</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>PASSMILES</td>
<td>Passenger miles</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>PASSHOURS</td>
<td>Passenger hours, including both runtime and dwell time</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>WAITHOURS</td>
<td>Passenger waiting hours, with wait calculated as 1/2 headway</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>PASSDELAY_DEP</td>
<td>Delay to passengers boarding at this stop (i.e. additional wait time)</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>PASSDELAY_ARR</td>
<td>Delay to passengers alighting at this stop (i.e. those late to their destination)</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>RDBRDNGS</td>
<td>Rear door boardings</td>
<td>Float</td>
<td>AVL-AKC</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>CAPACITY</td>
<td>Vehicle capacity</td>
<td>Float</td>
<td>AVL-AKC</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>DOORCYCLES</td>
<td>Number of times door opens and closes at this stop</td>
<td>Float</td>
<td>AVL-AKC</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>WHEELCHAIR</td>
<td>Number of wheelchairs boarding at this stop</td>
<td>Float</td>
<td>AVL-AKC</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>BIKERACK</td>
<td>Bikerack used at this stop</td>
<td>Float</td>
<td>AVL-AKC</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Category</td>
<td>Field</td>
<td>Description</td>
<td>Type</td>
<td>Source</td>
<td>Imputed</td>
<td>Weighted</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>---------</td>
<td>----------</td>
<td>---------</td>
<td>----------</td>
</tr>
<tr>
<td>Crowding</td>
<td>VC</td>
<td>Volume-capacity ratio</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>CROWDED</td>
<td>Volume exceeds capacity</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>CROWDHOURS</td>
<td>Passenger hours when volume exceeds capacity</td>
<td>Float</td>
<td>Calculated</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Additional ID Fields</td>
<td>ROUTE_ID</td>
<td>Route ID in GTFS</td>
<td>Integer</td>
<td>GTFS</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ROUTE_AVL</td>
<td>Route ID in AVL-APC</td>
<td>Integer</td>
<td>AVL-APC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TRIP_ID</td>
<td>Trip ID in GTFS</td>
<td>Integer</td>
<td>GTFS</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STOP_ID</td>
<td>Stop ID in GTFS</td>
<td>Integer</td>
<td>GTFS</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>STOP_AVL</td>
<td>Stop ID in AVL-APC</td>
<td>Float</td>
<td>AVL-APC</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BLOCK_ID</td>
<td>Block ID in GTFS</td>
<td>Integer</td>
<td>GTFS</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SHAPE_ID</td>
<td>Shape ID in GTFS</td>
<td>Integer</td>
<td>GTFS</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SHAPE_DIST</td>
<td>Distance Along Shape (m)</td>
<td>Float</td>
<td>GTFS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>VEHNO</td>
<td>Vehicle Number</td>
<td>Float</td>
<td>AVL-APC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. Sample Results

This section presents sample results from the data mashing tool. The values themselves have not yet been fully validated, so should not be treated as final statistics. Instead, the purpose of this section is to illustrate the types of performance measure the tool is capable of reporting, and how those measures might be useful in planning. In all cases, the performance reports seek to report information that is both relevant to the planning process and readily explainable to policy makers. It further seeks to put the focus of the analysis on the changes that occur over time, rather than a single snapshot of the system.

FIGURE 2 shows a sample of the monthly transit performance report. It consolidates the core performance measures onto a single page, and compares them to performance from another period, often the month before. The measures are grouped in the following categories:

- **Input Specification**: Attributes selected by the user to define the scope of the report. The geographic extent can be the bus system as a whole, a route or an individual stop, with some minor differences for the route or stop reports. The day-of-week is weekday, Saturday or Sunday/holiday. Time-of-day can be specified for the daily total, or for individual time periods allowing for evaluation of peak conditions. The report generation date and a comments section are provided. The notes in this case indicate that school returns to session between the two periods.

- **Service Provided**: The service provided metrics are simply measure of the total scheduled transit service, as found in the GTFS. Identical values mean that the schedule did not change between those two months.

- **Ridership**: Ridership measures provide the total passenger boardings, the distance and time passengers spend onboard, and the number of wheelchairs and bicycles served. In this example, the ridership increases by 6% from August to September, mostly likely because schools return to session.

- **Level-of-Service**: The level-of-service section provides measures of the quality of service provided, as experienced by users. The average run speed, the dwell time per stop and the headway are measured as a function of the busses themselves. Run speed is defined as the speed between stops, so excludes the dwell time at stops. Headway is measured at each route-stop based as the time from the previous trip of the same route. That is, it
does not account for combined headways across multiple routes. The fare is reported as the average full cash fare across all routes and stops, as shown in the GTFS. Separate revenue data would be needed to measure the average fare paid accounting for discounts and passes. The average distance traveled, average passenger speed and average passenger wait are measured as a function of the passengers themselves. In contrast to the run speed, the average passenger speed includes dwell time, making it generally slower. Average waiting time is measured as half the headway, assuming random passenger arrivals. Note that the system-wide average passenger wait tends to be less than half the system-wide average headway because passengers tend to use more frequent service.

- **Reliability**: Reliability measures indicate how well the busses adhere to their schedule. Two separate measures of on-time performance are reported, corresponding to short and long delays. About 60% of busses arrive within 2 minutes of their published schedule, and 90-95% arrive within 10 minutes of their published schedule. In addition, two measures of delay are reported which are weighted to passengers instead of busses. The waiting delay is the average time passengers wait at their stop for a bus to arrive after its scheduled arrival time. Arrival delay is the average time passengers arrive at their alighting stop, past the scheduled time.

- **Crowding**: The average volume-capacity ratio in this report is low because it is measured across all busses, whereas crowding tends to be an issue specifically in the peak period and in often in only one direction. Similarly, a low percentage of vehicles experience volumes in excess of capacity for the day as a whole, and a modest number of passenger hours occur in such conditions. In addition to the concentrated nature of these conditions, it is worth noting that a strict definition of crowding is used and that the capacity itself imposes a constraint on meeting that definition.

- **Observations**: The number of days and percent of trips observed are included to provide the analyst with a basic understanding of the limits of the data.

This performance report provides an overview allowing planners to quickly scan a range of indicators for changes that might be occurring.
## SFMTA Transit Performance Report

### Input Specification
- **Geographic Extent:** All Busses
- **Day-of-Week:** Average Weekday
- **Time-of-Day:** Daily Total
- **Report Generated on:** 13-Jul-2014
- **Comments:** School returns to session in September

### Periods

<table>
<thead>
<tr>
<th>Service Provided</th>
<th>Aug-09</th>
<th>Sep-09</th>
<th>Difference</th>
<th>% Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily Vehicle Trips</td>
<td>30,226</td>
<td>30,226</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td>Service Miles</td>
<td>149,401</td>
<td>149,401</td>
<td>0</td>
<td>0.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ridership</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Boardings</td>
<td>509,868</td>
<td>540,259</td>
<td>30,390</td>
<td>6.0%</td>
</tr>
<tr>
<td>Passenger Miles</td>
<td>1,004,107</td>
<td>1,057,421</td>
<td>53,314</td>
<td>5.3%</td>
</tr>
<tr>
<td>Passenger Hours</td>
<td>110,455</td>
<td>115,579</td>
<td>5,124</td>
<td>4.6%</td>
</tr>
<tr>
<td>Wheelchairs Served</td>
<td>1,067</td>
<td>1,099</td>
<td>33</td>
<td>3.1%</td>
</tr>
<tr>
<td>Bicycles Served</td>
<td>1,729</td>
<td>1,837</td>
<td>109</td>
<td>6.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level-of-Service</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Run Speed (mph)</td>
<td>11.95</td>
<td>11.93</td>
<td>-0.02</td>
<td>-0.2%</td>
</tr>
<tr>
<td>Average Dwell Time per Stop (min)</td>
<td>0.20</td>
<td>0.20</td>
<td>0.00</td>
<td>2.3%</td>
</tr>
<tr>
<td>Average Headway (min)</td>
<td>17.3</td>
<td>17.3</td>
<td>0.00</td>
<td>0.0%</td>
</tr>
<tr>
<td>Average Full Fare ($)</td>
<td>$2.06</td>
<td>$2.06</td>
<td>$0.00</td>
<td>0.0%</td>
</tr>
<tr>
<td>Average Distance Traveled per passenger (mi)</td>
<td>1.97</td>
<td>1.96</td>
<td>-0.01</td>
<td>-0.6%</td>
</tr>
<tr>
<td>Average Passenger Speed (mph)</td>
<td>9.09</td>
<td>9.15</td>
<td>0.06</td>
<td>0.6%</td>
</tr>
<tr>
<td>Average Passenger Wait (min)</td>
<td>5.56</td>
<td>5.54</td>
<td>-0.02</td>
<td>-0.3%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reliability</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent of Vehicles Arriving within 2 Minutes of Schedule</td>
<td>59%</td>
<td>56%</td>
<td>-0.03</td>
<td>-4.4%</td>
</tr>
<tr>
<td>Percent of Vehicles Arriving within 10 Minutes of Schedule</td>
<td>94%</td>
<td>93%</td>
<td>-0.01</td>
<td>-1.4%</td>
</tr>
<tr>
<td>Average waiting delay per passenger (min)</td>
<td>3.17</td>
<td>3.31</td>
<td>0.13</td>
<td>4.2%</td>
</tr>
<tr>
<td>Average arrival delay per passenger (min)</td>
<td>3.65</td>
<td>3.83</td>
<td>0.18</td>
<td>5.0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Crowding</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Volume-Capacity Ratio</td>
<td>0.23</td>
<td>0.25</td>
<td>0.01</td>
<td>6.3%</td>
</tr>
<tr>
<td>Percent of Vehicles with Volume &gt; Capacity</td>
<td>0.3%</td>
<td>0.3%</td>
<td>0.1%</td>
<td>31.4%</td>
</tr>
<tr>
<td>Passenger Hours in Crowded Conditions (V&gt;C)</td>
<td>2,078</td>
<td>2,281</td>
<td>203</td>
<td>9.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Observations</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Days</td>
<td>21</td>
<td>22</td>
<td>1</td>
<td>4.8%</td>
</tr>
<tr>
<td>Percent of Trips Observed</td>
<td>15.1%</td>
<td>14.1%</td>
<td>-1.0%</td>
<td>-6.8%</td>
</tr>
</tbody>
</table>

FIGURE 2 Sample Transit Performance Summary Report
While the numeric performance measures provide valuable information, their aggregate nature can wash out change that may be occurring in one portion of the city. Therefore, an interactive mapping tool was developed to plot key metrics in their geographic context. FIGURE 3 shows a screenshot from this tool. The left map shows a before period, the middle map an after period, and the right map shows either the absolute or relative change between the two periods. The user can select which time-of-day, which performance measure and which direction to plot. In this instance, the user has chosen to map the degree of crowdedness in the outbound direction during the 7-10 pm time period. The warm colors on the left two maps indicate more crowding, as measured by the average volume-capacity ratio during the period. The results are logical, with reasonably full busses moving west from the central business district towards residential areas of the city, as well as northwest on Columbus Avenue. The map on the right shows the relative change in the metric between the two periods, with the warm colors indicating an increase in crowdedness and the cool colors indicating a decrease. In this instance, crowdedness increases throughout much of the city during this period, due to some service reductions, but those areas showing change tend not to be very crowded to start.

To accommodate further analysis of the changes that occur to specific routes, the software generates route profiles as shown in FIGURE 4. In this example, average weekday ridership on the 1-California route is plotted in the inbound direction during the AM peak. The x-axis is the sequence of stops along the route. The line charts, associated with the right axis, show the number of passengers on the bus between each stop. The bar charts show the number of passengers boarding and alighting at each stop, with positive bars indicating boardings and negative bars indicating alightings. In all cases, the orange and red colors indicate the March 2010 period, and the blue colors indicate the June 2010 period. The chart shows higher overall volumes on this route during the AM peak time period in June compared to March. The pattern of ridership remains similar between the two periods, with riders accumulating through the residential portions of the route, and passengers getting off the bus when it reaches the central business district, starting at the Clay Street & Stockton Street stop. The PM peak ridership profile would show the reverse. These boarding profiles are useful when evaluating service changes made to specific routes, or the ridership resulting from newly opened land developments.
Finally, line plots are output, as in FIGURE 5, to show the trends over a longer period of time, rather than just for two periods. In this particular example, the average headway is plotted for six time periods throughout the day. The start of the time period is indicated by the numbers in the legend, in military time. The chart shows that during this three year period, the service provided is generally very stable. One notable exception to this stability occurs between March and September 2010. During this time period, bus service was cut by about 10% due to budget constraints, with the cuts concentrated in the late night (2200) time period. The service was partially restored in September 2010. This change shows up in the data as a “bubble” in the top line where the headway increases, and much more modest bubbles in the lower lines. Any of the performance measures can be easily plotted in this way, and doing so is an important step to understanding whether the changes observed are real, or simply within the natural variation of the data.

The software can automatically generate each of the performance reports described above, allowing for core analysis of the most important measures. In addition, the full weighted and imputed Datastore is available for advanced users who seek to conduct further in-depth analysis or custom queries.
FIGURE 3 Sample Transit Performance Change Map
FIGURE 4 Sample Route Profile
FIGURE 5 Sample Trend Plot

System-wide change in bus frequency by time of day

Minutes between buses
5. Conclusions and Future Development

The product of this research is a Big Data mashing tool that can be used to measure transit system performance over time. The software is implemented for San Francisco, but can be adapted for use in other regions with similar data.

The paper addressed some of the methodological and mechanical challenges faced in managing these large data sets and translating them into meaningful planning information. One such challenge was the sampled nature of the data, where not all vehicles have AVL-APC equipment installed. To make these data more representative of the system as a whole, the vehicle trips in the AVL-APC data are expanded to match the universe of vehicle trips identified by the GTFS data, missing data are imputed where possible, and weights are developed to scale up to compensate for data that cannot be imputed. The expansion process applies strategies from traditional surveys where a small but rich data set is expanded to match a less rich but more complete data set. Such strategies are key to spreading the use of Big Data for urban analysis beyond the first tier of cities that have near-complete data sets to those that are constrained by partial or incomplete data.

The software is available under an open-source license from (Erhardt, 2014). For working with these large data sets, it was an important decision to work with libraries that allow fast querying of on-disk data, but also the ability to easily modify the data structure.

The data mashing tool reports and tracks transit system performance in the core dimensions of: service provided, ridership, level-of-service, reliability and crowding. The performance measures are reported for the system, by route and by stop, can also be mapped using an interactive tool. The focus of the tool is on providing the ability to monitor the trends and changes over time, as opposed to simply analyzing current operations. By making performance reports readily available at varying levels of resolution, and the data mashing tool encourages planners to engage in data-driven analysis on an ongoing basis.

Several extensions of this research are currently planned. First, the outputs will be validated by comparing to other available measures, such as the officially published ridership totals. Second, if access can be negotiated to the transit smartcard data set, it will be incorporated into the process by expanding the sampled smartcard data to align with the APC
totals. Doing so would provide additional value by allowing transfers and linked trips to be monitored. Third, plans are in place to incorporate highway performance measures into a combined tool, allowing both to be tracked in concert.

Ultimately, the data mashing tool will be applied to measure the change in performance before and after changes to the transportation system. The study period covers a time with important changes to the transit system, such as a 10 percent service cut in 2010 implemented due to budget constraints (Gordon, Cabanatuan, & Chronicle Staff Writers, 2010), and several pilot studies aimed at improving the speed and reliability of transit service in specific corridors (City and County of San Francisco Planning Department, 2013). Evaluating these changes will provide planners and researchers with greater insight into the effects of transportation planning decisions.

Acknowledgements

The authors would like to thank the San Francisco County Transportation Authority (SFCTA) for funding this research, the San Francisco Municipal Transportation Agency (SFMTA) for providing data, and both for providing valuable input and advice.


Guiding Data-Driven Transportation Decisions

Kristin A. Tufte
Departments of Computer Science and Civil and Environmental Engineering
Portland State University
P.O. Box 751
Portland, OR 97207-0751
Email: tufte@pdx.edu

Basem Elazzabi
Department of Computer Science
Portland State University
Email: elazzabi@pdx.edu

Nathan Hall
Department of Computer Science
Portland State University
Email: nahall@pdx.edu

Morgan Harvey
Department of Civil and Environmental Engineering
Portland State University
Email: morgan@pdx.edu

Kath Knobe
Intel Corporation
Email: kath.knobe@intel.com

David Maier
Department of Computer Science
Portland State University
Email: maier@pdx.edu

Veronika Megler
Department of Computer Science
Portland State University
Email: vmegler@pdx.edu
Guiding Data-driven Transportation Decisions

ABSTRACT
Urban transportation professionals are under increasing pressure to perform data-driven decision making and to provide data-driven performance metrics. This pressure comes from sources including the federal government and is driven, in part, by the increased volume and variety of transportation data available. This sudden increase of data is partially a result of improved technology for sensors and mobile devices as well as reduced device and storage costs. However, using this proliferation of data for decisions and performance metrics is proving to be difficult. In this paper, we describe a proposed structure for a system to support data-driven decision making. A primary goal of this system is improving the use of human time, effort and attention with side benefits of improved consistency and documentation.

INTRODUCTION
Data-driven decision making and data-driven performance metrics are a high priority in the transportation world today and are the focus of much attention and effort. An influx of new sensors and data from those sensors along with pressure from the federal government and new federal regulations, such as MAP-21 (FHWA, 2102) is driving the focus on data-driven decisions and metrics. However, turning this explosion of new data into actionable information is difficult. The current state of support in data-driven performance reporting is limited by several factors including data stored across disparate locations and systems, insufficient documentation of how data was chosen and assembled, and manual manipulation of the data. These factors limit the ability of urban transportation professionals to produce and reuse the work done to create data-driven performance metrics.

In this paper, we present a proposed system structure that begins to address these issues. Our system aims to support integration across diverse-distribution data sources and to increase re-use of the data and processing mechanisms. We aim to improve productivity by improving the effective use of human time. Our system contains three key parts: The Portland Observatory, Guides and Concurrent Collections (CnC) (Burke, Knobe, et al. 2010). The Portland Observatory collects urban transportation-related data in a single location (or common portal), Guides
represent decisions made in selecting and assembling data related to a decision or performance metric and CnC helps automate Guide instantiation.

As a simple example of the motivation for our system, consider creating a mobility performance report. A mobility report is a common type of report generated by transportation professionals; such a report describes the ability of the population to move around an urban area and uses metrics such as travel time, travel speed and reliability of travel time. Creating such a report requires first tracking down and selecting data sources followed by data cleaning and analysis and generation of graphics. We observe that decision-making tasks are recurrent - similar decisions may need to be made for reports created for different locations and time periods (i.e. annual reports).

The first step in creation of this report is tracking down data sources. In the current state these data sources may be on local hard drives, on central storage servers at possibly different agencies or in other disparate places. The Portland Observatory collects data in one location or portal simplifying the process of tracking down data sources. A second step in creating the report is to perform analysis to produce mobility performance metrics; this step may include data cleaning, analysis and generation of graphics. We observe that decisions, such as excluding a set of sensors due to construction, are typically made along the way. A Guide is like a template for a decision-making task; Guides capture the input data criteria, data cleaning and analysis processes and decisions. By capturing the cleaning, analysis and decision process, Guides enable retrospective review of the performance metric-generation process. In addition, Guides enable reuse of the work done generating performance metrics and help ensure that analysis is consistent across decisions and reports. The CnC parallel programming environment is used to automate Guide instantiation to improve re-use and productivity.

We begin this paper by presenting three example reports produced in the Portland, OR-Vancouver, WA metropolitan region. We discuss each report and then summarize the connections and similarities between the reports and how those similarities might be leveraged. We next present Guides - our technique for encapsulating the report-creation process and context - we discuss guide motivation, content and structure. We proceed to discuss data collection and
opportunities for leverage therein. Finally, we describe a proposed system architecture and conclude with a description of our prototype guide implementation.

EXAMPLE REPORTS
We begin by describing three reports that are currently generated in the Portland, OR-Vancouver, WA metropolitan region: The Metro Performance Measures Report, the RTC Congestion Management Process Monitoring Report and the Metro Portal Annual Report. For each report, we identify a key product in the report that we will discuss in the comparison section below and use in our guide prototype implementation.

Metro Performance Measures Report (Portland, OR)
The Metro Performance Measures Report is a performance report that is generated every two years by Metro, the regional government for the Portland, OR region. The report is generated in response to Oregon State Statues and consists of 12 measures related to land use planning and coordination including development density, job creation, land use and transportation measures. Measure 9 in the report contains transportation-related measures - specifically "Transportation measures including mobility, accessibility, and air quality indicators" (Metro, 2011). We focus on the mobility and accessibility metrics, which include vehicle miles traveled, travel time reliability for major freeways, and transit ridership. The vehicle miles traveled in the Metro report is taken from the FHWA State Performance Monitoring system. The travel time reliability numbers come from the data in the Portal data archive. Portal is the regional transportation data archive for the Portland, OR-Vancouver, WA metropolitan region (Portal, 2014). The key product in the Metro Performance Measures Report that we use for analysis and prototyping is a table of travel time reliability measures that includes average travel time and average congested travel time for freeways in the Portland region; Table 9.2 in the 2011 report.

**Key product:** Table of travel time reliability.

RTC Congestion Management Process Monitoring Report (Vancouver, WA)
The Congestion Management Process Monitoring Report is an annual report produced by the Southwest Washington Regional Transportation Council (RTC). Federal law requires that RTC
maintain a Congestion Management Process (CMP); the CMP Monitoring report is part of RTC's CMP process. The annual report "provides a comprehensive set of data for monitoring the performance of the transportation system" and "provides on the travel characteristics of the regional transportation corridors" (RTC, 2013). The report contains a System Monitoring section that provides system performance measures including vehicle volumes, capacity ratio, travel speed and intersection delay, vehicle occupancy, safety, truck percentage and several transit measures including ridership and on-time performance. The key product in this report that we focus on for analysis and prototyping is the Speed: Auto Travel Speed and Speed: Speed as Percent of Speed Limit metrics. These speed metrics are presented as maps in the report; specifically, maps 8, 9, 10 and 11 in the 2013 report. We note that at the current time, the speed metrics are based on probe runs. However, once confidence in the data is established, these metrics could be based on data collected through automated sensors, such as Bluetooth detectors and high-definition radar. Basing the metrics on installed sensors has the potential to increase temporal and spatial coverage of the metrics. Though not discussed in this paper, the report also uses Vehicle Volume metrics.

**Key product:** Maps of Auto Travel Speed and Speed as a Percent of Speed Limit

**Metro Portal Annual Report (Portland, OR)**

The Metro Portal Annual Report is a small report produced annually by the Portal team at Portland State University. The report was first produced in 2012 and is expected to be generated annually. This report contains two primary types of products: maps of speed and plots of volume by highway. Figure 1 shows a map and a sample plot from the 2012 Portal Annual Report. The map shows average vehicle travel speeds in the PM Peak period (4-6PM) for the freeways in Portland in 2012. The plot shows 15-minute volumes by time of day for mid-weekdays for the I-5 freeway in Portland at Portland Boulevard. The report contains a series of such plots for key freeway locations.

**Key products:** Travel Speed Maps and Volume Plots by Highway.
Report Comparison and Observations

A key observation is that the key products identified above for the three reports all use vehicle speed data as input. One report presents travel time, the other two present speed. Nevertheless, speed is a base data source (We note that travel time can be calculated as segment length divided by speed.) Given that all three reports use speed as an input data source, all could potentially draw from the same data source. The Guide Instantiation section near the end of this paper shows how Guides capture the fact that speed data is an input to the three products and show how the three products all use the same input Guide. In addition to using the speed data as input, several of the reports make selections such as using data for a defined peak period and using only data for mid-week days. Consistency in these types of selections and in data cleaning and metric calculation algorithms across these reports is likely to be beneficial, particularly for reports generated for the same region - such as the two reports generated for the Portland metropolitan region. At a minimum, the ability to be consistent - that is the ability to use the same peak period definition, the same mid-weekday selection and the same metric calculation algorithms, if desired, would be very useful. In creating the Portal annual report, the Portal team intentionally
used the same selections as in the Metro Performance Measures Report. Capturing those selections in a formal process would be useful for future report creation. As we will describe, Guides also capture these types of selections.

GUIDES
In one sentence, a Guide is an encapsulation of the experience required to answer a question. We proceed to discuss Guide motivation, concept and structure.

Guide Motivation
The reports described above currently require manual effort for data collection and combination and review and cleaning. In some cases, the process is retained in human memory and is not documented. The nuanced decisions about data cleaning and combination, which come about after years of producing reports are difficult to articulate and document. However, we observe that some pieces of the report creation can be automated. Guides are designed to capture the pieces of report generation that can be captured.

One might first consider some alternatives. Data integration has been an area of research in computer science for years (Halevy, 2006). However, a heavy-grade data integration approach may not be justified for producing transportation performance reports - the example reports are generated annually or bi-annually, for this type of reporting, even for quarterly reporting, the cost of maintaining an integrated schema is too much; however, making the data sources accessible through a common portal is useful. Creating a common portal could mean putting the data in a database where it can be accessed with SQL queries. SQL queries are very useful and powerful, but they don't record judgments or the reasoning behind the judgments made in report creation. In short, traditional data management techniques may be overkill (i.e. data integration) or may not capture the required information (i.e. SQL queries).

While we understand that we can't totally eliminate human involvement, we do believe that we can we make it easier for humans to be involved, increase the efficiency of human involvement and capture some of the human judgments - i.e. the reason a set of data was eliminated from consideration.
Guide - Concept

A Guide is a set of information that describes and encapsulates a decision-making process. By describing and encapsulating a decision-making process, a guide may enable reuse of data or algorithms by making it easier to switch data sources, to respond to a change in the format of a data source or to re-create a report in a different context - i.e. take a report done for one metropolitan area and create it for another metropolitan area.

We identify the following objectives for guides:

- A guide should capture both the data requirement and how the data requirement is being met in a particular instantiation of the guide.
- A guide should embody requirements and expectations as well as results.
- A guide should also capture weaknesses of the current data set so that if the ideal data set becomes available or is available in a different locality, it can be used (i.e. volumes by vehicle type were the desired data source, but that data was not available).
- Guides should capture requirements, not just process. The process may over- or under-constrain due to specifics of the locality and time for which the report is generated. The process may capture which data is not used for the report, it may not capture why that data was not used.

Guide Structure

We have identified the following fields as being part of a Guide.

1. Question - The question answered by a Guide.
2. Name - The guide name.
3. Description - A text description of the guide.
4. Parameters - Parameters to the guide such as spatial location or time period.
5. Input Data - A list of data sources and variables (attributes) needed from each data source.
6. Queries and Analysis - Code such as SQL queries for analyzing the data.
7. Output Data and Format - A list of variables (including types) in the output, output format and output structure (or organization).
9. Quality - Some indication of the quality of the results such as a list of suitable uses.
10. Comments - Things you learned along the way.

Guide Creation and Re-Use
Another motivation for guides is to leverage expert knowledge. A particular analyst may be an expert in travel time calculation algorithms or a particular MPO (Metropolitan Planning Organization) may create a report to meet a specific regulation. Guides are intended to capture such knowledge so it can be easily transferred to other analysts or MPOs. We imagine an initial guide being created by highly skilled person who selects data and specifies computations and builds the guide from scratch. A medium skilled person may be able to take that guide and customize it for a time period or a metropolitan area; such a person may understand the general concept of the report, but may not be familiar with the details of the calculation algorithms. Finally, we imagine a lower skills person may be able to simply re-run a report by specifying parameters such as a desired year. This person may not have an understanding of the report, but is able to create different versions of the report.

DATA COLLECTION
As discussed, a key issue in creating performance metric reports is locating and combining data. The Portland Observatory acts as an aggregator and repository for transportation data acquired from a variety of agencies in the Portland-Vancouver metropolitan area. The Portland Observatory leverages Portal, the Portland-Vancouver region's transportation data archive (Portal, 2014). Portal consists of a PostgreSQL database containing approximately 3TB of transportation-related data collected from Portland-Vancouver area agencies over the past ten years and a web interface. The data in Portal includes freeway loop detector data, weather data, Bluetooth travel time data, weigh-in-motion data, transit data, arterial signal data and more. Figure 2 shows two screenshots from the Portal Data Archive web interface.
Figure 2 Portal Data Archive Web Interface

Data Collection

Data flows into Portal from a variety of sources and systems, collected by agencies using a variety of sensor technologies and vendor systems supporting them. Table 1 shows the Arterial, Transit and State Highway data sources for the Portal archive by agency and type of data. In Table 1, sources in the process of being integrated are in italics. All other sources arrive in automated or semi-automated (in a few cases) fashion. Considering Table 1, we observe a few patterns of interest from a data management perspective.

Common Data Formats: Some types of data are produced by multiple agencies. Both Transit agencies - TriMet (OR) and C-TRAN (WA) produce GTFS data. In this case, the data produced is in identical formats across systems.

Common Systems: Several agencies use common systems. For example, agencies that manage traffic signals in the Portland area use the TransSuite central system, agencies that manage traffic signals in Vancouver use the ATMS.Now central signal system.
Common Types of Data: Many agencies collect similar types of data, but from different vendors and different sensors. For example, travel times from Bluetooth readers are collected by ODOT, City of Portland and Clark County. The detection technology is similar, but the devices and software vendors are different for all three agencies.

Table 1 Portal Archive Data Sources By Agency

<table>
<thead>
<tr>
<th>Arterial</th>
<th>Transit</th>
<th>State Highways</th>
</tr>
</thead>
<tbody>
<tr>
<td>City of Portland (OR)</td>
<td>TriMet (OR)</td>
<td>ODOT (OR)</td>
</tr>
<tr>
<td>- TransSuite Central Signal System</td>
<td>- AVL/APC</td>
<td>- Loop Detectors</td>
</tr>
<tr>
<td>- Travel Times</td>
<td>- GTFS</td>
<td>- High Definition Radar</td>
</tr>
<tr>
<td>Washington County (OR)</td>
<td>C-TRAN (WA)</td>
<td>WSDOT (WA)</td>
</tr>
<tr>
<td>- TransSuite Central Signal System</td>
<td>- AVL/APC</td>
<td>- Loop Detectors</td>
</tr>
<tr>
<td>Clarkamas County (OR)</td>
<td></td>
<td>- High Definition Radar</td>
</tr>
<tr>
<td>- TransSuite Central Signal System</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clark County (WA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- High-Definition Radar</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- ATMS.Now Central Signal System</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- Travel Times</td>
<td></td>
<td></td>
</tr>
<tr>
<td>City of Vancouver (WA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>- ATMS.Now Central Signal System</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Opportunities for Leverage

We discuss each of the observations and the opportunities for leverage it provides.

Common Data Formats - Opportunities for Leverage: When data is provided to an archive, such as Portal or the Portland Observatory, from multiple sources in a common format, the integration of data from those sources is straightforward. In addition, any performance metrics and visualizations such as plots or maps made on one of the data sources can typically be easily
extended to data from the other data sources. Google has developed the General Transit Feed Specification (GTFS) for describing transit schedule and arrival data (Google, 2014). This format greatly simplifies processing and visualizing transit data across agencies. GTFS is one of the few commonly used Common Data Formats of which we are aware.

Common Systems - Opportunities for Leverage: Both the Portland, OR and Vancouver, WA regions use common central signal systems; though the signal system used by the Portland region is different than the signal system used by the Vancouver region. Data coming from a common system provides similar and additional benefits in terms of leverage compared to data with Common Data Formats. Having the data come from a Common System, we assume the data format will be the same. In addition, having data come from the same system (either a single or multiple installations of that system) can leverage transfer mechanisms or protocols established for that system.

Common Types of Data - Opportunities for Leverage: The phrase "Common Types of Data" refers to data that is the same "type" of data, but comes from different systems and in different formats. Extending from the Common Systems discussion above; we comment that signal data from the Portland signal system and signal data from the Vancouver signal system qualify as "Common Types of Data" – the data from both systems represents signal operations, but comes in different formats and requires different network transfer mechanisms. Bluetooth data is another example of this type of data in the Portland-Vancouver region. Bluetooth data is provided to the Portal archive in three different formats from three systems. Common Types of Data do provide opportunities for leverage; however, leveraging such data requires additional work. Common features of the data that occur across the different data sources must be identified. Products can then be built on the identified common features. In our experience, this type of data collection - Common Types of Data – where data is similar, but without a common format and not from a common system occurs regularly.

It is a goal of the Portland Observatory to address this issue of integrating across "Common Types of Data." A user producing a performance report does not care what format the data arrived in or what system the data is from (except, perhaps as those relate to data quality). One
purpose of the "usifications" in the Portland Observatory, described below, is to help abstract away the incoming data format and system-specific information, so a general user does not need to worry about those issues and can focus on processing the data itself. As discussed before, there is a large volume of work in the Computer Science literature on data integration; however, we believe that work provides a heavyweight solution, when a lightweight solution is more appropriate.

**PROPOSED SYSTEM ARCHITECTURE**

We propose an architecture for the Portland Observatory to process and "usify" data. When the data arrives at the Portland Observatory, we propose to store it in the Portal “raw” form, that is, in the form that it arrived from the source. Some of the raw data is then “usified”; that is, cleaned, aggregated, added to or otherwise transformed into a form more directly usable by observatory clients. For example, data from multiple sensor types and provided by multiple agencies may be combined to provide traffic speeds and counts across the metropolitan Portland area. The same raw data may be transformed in different ways for different users, for example aggregating to different intervals or applying different cleaning methods. Keeping the raw data allows researchers and analysts to test different methods or models; improvements can be implemented as new “usifications”.

In the Portland Observatory, developers create Guides that select, combine, aggregate or subset the data for the users. As discussed, each Guide encapsulates a specification for and produces a set of data to meet a specific need and a Guide can reuse another Guide as a component, possibly specifying some different parameters; this reuse provides consistency and allows for faster development. For example, a Guide for Portland, OR can be reused by Vancouver, WA by changing the geographic boundaries. If a new cleaning method or model becomes available via a new usification, a Guide can be altered to use it; all reusing Guides can immediately reap the benefits, without needed to be altered themselves.

Intel Concurrent Collections (known as CnC) is used to help automate Guide re-use (Burke, Knobe, et al. 2010). CnC is a programming model that supports a declarative description of an application (the analysis portion of a Guide in our case). Typically applications are represented
Guiding Data-Driven Transportation Decisions

programmatically; that is, an application consists of files of computer code. In contrast, CnC's application specification is based on the idea of whiteboard drawings. The functionality of an application may be drawn as a graph on a whiteboard. To use CnC, the drawing of the graph is translated to a text-based graph specification in the CnC language; then the CnC system executes the application based solely on the CnC graph specification. To verify what a traditional application is doing, one must examine the computer code. (The computer code may be accompanied by a description or specification; but examining the description or specification does not verify functionality of the code.) In contrast, to verify what a CnC program is doing, one inspects the graph-based CnC specification. The CnC specification is typically much easier to read and understand than computer code. This ability to declaratively describe an application is the key feature of CnC with respect to the Portland Observatory.

![Proposed System Architecture Diagram](image)

**Figure 3 Proposed System Architecture Diagram**

Figure 3 shows a diagram of the proposed architecture of the system. As discussed above, our system combines a relational database system (RDBMS) and the CnC parallel programming environment. The user application is written with CnC and accesses data stored in the relational database. Combining a parallel programming environment (such as CnC) with a relational database system is our approach to addressing the variety and varied data sets inherent in big data, and supports scalability when needed by large datasets. A transportation engineer trying to produce a report must integrate data from disparate data sources and must clean and process such data. As data sizes increase, it will be increasingly important for users to think about the application problem separately from the data layout and parallelization of the problem. Both CnC
and RDBMSs provide conceptual, high-level, declarative interfaces that allow this separation: the RDBMSs provide declarative access to the data, and CnC enables programmers to declaratively assemble their application from lower-level functions. Having the process and data accessed documented in a declarative language allows us to document how data has been combined - and, equally importantly for privacy and security, what data has not been accessed or combined in support of a decision. For example, this documentation can assist an agency in showing that they have acted in accordance with rules and policies governing them.

The technical support required to interface the systems is, in some ways, the easier part of the work. The more difficult and time-consuming part of these integrations is building the contacts and relationships with the various agencies, and gaining agreement to share and trust their data with an external aggregating party. Portland State’s experience in collaborating with these agencies over the past ten years provides a stable basis for these continuing and expanding collaborations.

PROTOTYPE IMPLEMENTATION

Using the proposed guide concept and structure and the system architecture described in previous sections, we implemented a prototype of the three key products identified in the Example Reports section: the travel time table, the travel speed maps and the volume plots. The prototype implementation has two parts: Guide Capture and Guide Instantiation. Guide Capture is a graphical interface that is used to capture features of a guide such as guide name, input data type, parameters, decisions and comments. The Guide Instantiation is a system that executes the analysis portion of a Guide and returns the requested results. We note that the Guide Capture and Guide Instantiation prototypes do not capture the full structure of a Guide as was described above; our prototype is in its early stages.

Guide Capture

The Guide Capture is a part of the Portland Observatory web interface. Figure 4 shows a screenshot of the main page of the Portland Observatory web interface. The web interface supports the display of a variety of data layers including air quality, weather, Bluetooth travel
times, bus stops, bicycle routes, tax lots and school sites. Some data in the observatory is
relatively static such as school sites, bicycle routes and parking meters.

![Portland Observatory Web Interface]

**Figure 4 Portland Observatory Web Interface**

The Portland Observatory interface supports a set of selectors to allow a user to display selected
data layers for selected dates and times. These selectors appear on the right-hand side of the
Portland Observatory web interface shown in Figure 4. The Guide Capture interface parallels and
leverages these selectors. To create a Guide, the user clicks on the "Create Guide" button in the
lower right-hand portion of the interface. This brings up a "Create Guide" dialog, shown in
Figure 5, that allows the user to specify the guide name and description and make selections
associated with the guide. The selectors in the "Create Guide" dialog (Layers, Year, Start Time)
are similar to those on the right-hand side of the web interface and the selectors in the "Create
Guide" dialog default to the current selections from the web interface. The motivation is for the
user to use the Portland Observatory interface to create a map of the data they want and then use
the "Create Guide" dialog button to capture those selections in a Guide.
A relevant feature of the Portland Observatory interface is the ability to add comments to data layers. Figure 6 shows the Portland Observatory comment interface. A comment can be added to a data layer at a specified location and time. The comments interface is currently used to record decisions and comments. Comments recorded for a data layer are associated with Guides created using for that layer.

Once a Guide is created, it needs to be instantiated with parameters and executed. We describe this process in the next section.
Guide Instantiation

In addition to a Guide Capture interface used to create guides, we have created a Guide Instantiation system and Guides for the key products identified in the Example Reports section. The three Guides for the three key products co-exist and operate in one centralized system. These Guides can communicate with each other, reuse each other, and produce different results depending on the request.

Guide Instantiation is made available as a service, which is used by the web interface. The Portland Observatory web interface receives a request from a user to instantiate a Guide and passes those requests to the Portland Observatory back end. The back end receives requests for Guide instantiation through the Guide Dispatcher in the form of a guide id and a set of parameters. For example, a guide for the Metro Portal Annual Report might have the year for which the report is to be generated as a parameter. The guide id tells the dispatcher what Guide should be instantiated. We note that guides are hierarchical and thus this request tells the dispatcher which Guide should be called to process the instantiation request, we call this Guide...
the "main Guide." When processing the request, the main Guide (the Guide that first received the request) can ask other Guides (sub-Guides) to provide sub-answers for the main Guide to construct or calculate the final result (answer). The sub-Guides themselves could request sub-answers from other Guides, and so on. In other words, the main request may generate a chain reaction inside the Portland Observatory across different Guides to make the final answer. The important thing to notice here is that, although each Guide has its own set of parameters, it is only the responsibility of the immediate caller to know what those parameters are and how they should be used; the main caller need only know the parameters for the Guide it calls; the main caller does not need to know about parameters for the Sub-Guides. Figure 7 shows the Guide Instantiation system.

![Figure 7 Guide Implementation Architecture Diagram](image)

Going back to the example reports, agencies may ask for data to generate these reports: travel times, speed, speed as a percentage of the speed limit, volume, and travel time for a specific year. Although each of those reports is different from an analyst’s perspective, the foundation of all of these reports is based on the same data using the same analysis methods. Thus, as a basis, the implementation uses a Guide that receives requests about speed. Giving the requirements determined by the parameters sent along with the request, the Guide knows what equations to use and on what data. To generate a report for the speed as a percentage of the speed limit, we only need two things, speed and speed limit. Since we already have a Guide that calculates speed, the Guide that generates the speed as a percentage of the speed limit can ask the speed Guide to
answer the speed question then combine the answer with the speed limit to get the final results. Generating the travel time report follows the same pattern as well since travel time requires speed and distance. Figure 8 shows the guide hierarchy for these reports.

![Figure 8 Guide Hierarchy Diagram](image)

**Figure 8 Guide Hierarchy Diagram**

An important thing to notice about Guides is that they don’t provide a complete solution; they only answer a question. Providing the complete solution is the end-application’s job. In other words, a Guide can be used by many, completely different applications that provide completely different solutions.

As a proof of concept, using our web interface, two types of results can be obtained using the Metro Portland Annual Report volume Guide: an Excel spreadsheet with transportation volume data for each highway along with their charts comparing volumes for different years and a map showing volume data for selected highways. Figure 9 shows a diagram of this result output.
CONCLUSION

We have presented a proposed system architecture and prototype implementation designed to support data-driven decision making, particularly in the transportation domain. The system uses Guides - an encapsulation of the experience required to answer a question. We selected three transportation system performance reports and key products to use examples for understanding and implementing Guides. Guides for these reports capture information such as the data required for the report, the parameters or selections needed for the report, the decisions made along the way and more. Based on our proposed system architecture, we implemented a prototype of Guides – front end Guide Capture using a web interface and back end Guide Instantiation using CnC and the PostgreSQL relational database. We call our integrated system the Portland Observatory. The system aims to increase integration across diverse data sources and to increase the re-usability of the transportation-related data and analysis that lies behind data-driven performance metrics and decision making and to make better use of valuable human time.
ACKNOWLEDGMENTS
This work was supported in part by the Intel Science and Technology Center for Big Data and a Maseeh Professorship. The authors acknowledge their collaborators in the transportation profession in the Portland-Vancouver region including, but certainly not limited to, Steve Callas (TriMet), Larry Ham (C-TRAN), Bob Hart (RTC), Rob Klug (Clark County), Peter Koonce (City of Portland), Deena Platman (DKS Associates), Stan Markuson (WSDOT), Dennis Mitchell (ODOT), Dale Robbins (RTC), Willie Rotich (City of Portland). This work would not be possible without the atmosphere of collaboration and cooperation that exists among the transportation professionals in Oregon and Southwest Washington.

REFERENCES
Burke, M., Knobe, K., Newton, R., Sarkar, V. The Concurrent Collections Programming Model. Technical Report TR 10-12, Department of Computer Science, Rice University, December 2010).


https://developers.google.com/transit/gtfs/


Computable Liveability

D. Glenn Geers · D. Economou

July 15, 2014

Abstract While there is no consensus on the definition of liveability, the term typically takes into account social, economic and environmental measures and is used widely, based on systems developed by the The Economist, PwC or other institutions. Given this lack of consensus, what does it mean when a city reaches the top of the ladder as ‘most liveable’? Is it possible to find a range of sometimes subjective quality of life indicators, consistent across the vast range of cities throughout the world: from Kabul to London, from Sao Paulo to New York, from Beijing to Sydney; and link them to objective measures and live city data? Can we find a single number that is meaningful and could, more importantly, be used as the basis for helping understand where best to invest resources in a city’s infrastructure? In the following we discuss some of the definitions of ‘liveability’ that have been adopted and examine their generality. For example, the three most widely quoted indices (Economist, Monocle and Mercer) may be more applicable to Western cities. The Economist includes measures of church attendance and union membership in its Index which may be more or less relevant depending on culture. Next we introduce the concept of Computable Liveability with the aim predicting liveability metrics based on objective measures alone. The Economist’s Index is a good example to study because they release information about their processes and, in particular, they show how they combine nine objective measures with survey results. Using non-parametric statistical analytics techniques we propose researching whether it is possible to learn the connection between subjective survey results and hundreds of objective measures and live urban data, and from this construct a potentially more general and consistent liveability index with usable predictive ability. Such relationships could be learned for a single city using historical data over time, or across different cities, or across different cities over time. Once a ‘Computable Liveability’ model has been learned from a consistent set
of urban data (taken from a range of cities) it should also be possible to generalise to non-
surveyed cities based on measured city data or predict future liveability scores for a single
city. Further, as the live data varies in real-time it may also be possible to compute real-time
liveability. Such methods would also be applicable to individual quality-of-life indicators.
Finally, we will outline a research program for realising our goal. Ultimately, this will con-
sist of constructing and delivering a liveability survey formulated in conjunction with social
scientists in a range of cities (but could in the first instance take the results from an existing
survey) and identifying the consistent set of cross-city data needed to generate our data-
driven liveability model. We will then use standard machine leaning methods to validate our
results. To conclude the study, we will compute indices for previously unsurveyed cities and
compare our results. From a practical point of view it may be more attractive to first perform
the work in a single country (or city) due to data availability, time and cost constraints. This
would quickly reveal both the usefulness of our proposed method and identify any issues.

Keywords Liveability ∙ Machine Learning ∙ Data driven modelling

1 Introduction

The idea of rendering the operations, machinations and human expectations of a city into a
single number termed ‘liveability’ appears to be popular ([9], [6] and [1] are just samples
from the press). Every year or so different organisations release ranked lists of liveability for
cities throughout the world. These rankings are reported in the media—‘City X is the most
liveable’; ‘City Y has fallen in the rankings’—they are used for tourism campaigns and for
encouraging businesses to relocate and generally as part of branding for major urban centres
competing globally. In this paper we will present the notion of ‘Computable Liveability’, the
ultimate aim of which is to be an additional policy assessment tool for City authorities. The
key idea is to realise principled mappings between objective measured data and existing
subjective survey (or other) results, with enough reliability that the tool can be used as
input into important decisions such as where best to invest in a city’s infrastructure for
example, with limited funds would a new school or a new road lead to improved liveability?
In so doing we will also examine the methodologies adopted in computing some of the
more widely reported liveability indices to see how they can inform our machine learning
approaches. This leads to a second goal of introducing some of the newer machine learning
techniques to the social sciences domain.

We distinguish between the terms ‘Liveability’ and ‘Quality-of-life’, using the former to
refer to a more subjective measure of city (or urban) performance (as collected by a survey
for example) and the latter for a more objective measure. This usage is in agreement with
the new ISO standard for City Indicators [20].

2 Liveability

There are a number of widely quoted Liveability indices that are reported more or less
annually. In this section we discuss three of the better known and more widely reported
indices before defining and discussing ‘Computable Liveability’ which forms the basis for
the rest of the paper.
2.1 A Discussion of some liveability indices

While we do not define particular liveability metrics, we will discuss the general methodology used to formulate three widely reported liveability indices: those from Mercer, Price Waterhouse Coopers (PWC) and The Economist (EIU). It is also worthwhile noting that these indices do not include measures of uncertainty, and that the machine learning techniques we propose to introduce can also provide uncertainty estimates. Table 1 shows the top ten cities for the three indices (in 2012) we discuss in the following subsections. While they measure similar properties there are striking differences between the rankings.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Mercer (Dec 2012 [7])</th>
<th>PWC (Oct 2012 [29])</th>
<th>EIU Survey (Aug 2012 [8])</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Vienna</td>
<td>New York</td>
<td>Melbourne</td>
</tr>
<tr>
<td>2</td>
<td>Zurich</td>
<td>London</td>
<td>Vienna</td>
</tr>
<tr>
<td>3</td>
<td>Auckland</td>
<td>Toronto</td>
<td>Vancouver</td>
</tr>
<tr>
<td>4</td>
<td>Munich</td>
<td>Paris</td>
<td>Toronto</td>
</tr>
<tr>
<td>5</td>
<td>Vancouver</td>
<td>Stockholm</td>
<td>Calgary</td>
</tr>
<tr>
<td>6</td>
<td>Dusseldorf</td>
<td>San Francisco</td>
<td>Adelaide</td>
</tr>
<tr>
<td>7</td>
<td>Frankfurt</td>
<td>Singapore</td>
<td>Sydney</td>
</tr>
<tr>
<td>8</td>
<td>Geneva</td>
<td>Hong Kong</td>
<td>Helsinki</td>
</tr>
<tr>
<td>9</td>
<td>Copenhagen</td>
<td>Chicago</td>
<td>Perth</td>
</tr>
<tr>
<td>10</td>
<td>Bern</td>
<td>Tokyo</td>
<td>Auckland</td>
</tr>
</tbody>
</table>

Table 1 Top Ten Cities in 2012 for the indices discussed below

At a purely formal level a liveability index, \( L \), may be calculated as a linear combination of quality-of-life indicator scores as shown in Equation 1,

\[
L = \sum_{i=1}^{N} w_i Q_i
\]

where the \( Q_i \) are quality-of-life indicator scores that may be objective or subjective measures and the \( w_i \) are weights which may be derived empirically or by other means. Normalization for sample size if required, is built into the weights. The three liveability measures described below subscribe to this scheme.

2.1.1 The Mercer Quality of Living Survey

The Mercer Index [24] is a purely survey based index which ranks 223 cities worldwide. The survey is primarily designed to help HR Managers plan for the relocation of staff.

The survey comprises thirty-nine questions covering ten quality-of-living factors: Political and Social Environment, Economic Environment, Socio/Cultural Environment, Medical and Health Considerations, Schools and Education, Public Service and Transportation, Recreation, Consumer Goods, Housing and Natural Environment. The survey questions, which do not take into consideration national or cultural factors, are scored on an eleven level Lickert scale (from zero to ten). Scores are weighted based on which quality-of-living factor they measure. The weightings were calculated from a pilot study which questioned responders on the importance of each of the ten quality-of-living factors. Political and Social Environment is weighted most highly while Schools and Education are rated least highly. That Schools and Education has the lowest weighting may be due to the fact that families
with school age children are less mobile than those without, but this would require further
investigation.

The number of people surveyed is not reported making it impossible to place error
bounds on the liveability scores, making the rankings unverifiable.

2.1.2 Price Waterhouse Coopers Cities of Opportunities

The Price Waterhouse Coopers Cities of Opportunity [28] report looks at a number cities
(thirty in the sixth—2014—edition) around the world and scores them in ten areas: Intel-
tlectual capital and innovation; Technology readiness; City gateway; Transportation and
Infrastructure; Health, safety and security; Sustainability and the natural environment; De-
mographics and liveability; Economic clout; Ease of doing business and Cost. Notice that
‘Demographics and liveability’ is merely one area in this work. However, we note that each
of the other nine areas tend to fall into the components which other reports use to form a
complete score, hence we treat the overall score—which is the un-weighted sum of the area
scores—as the liveability score.

PWC ‘upgrade, enhance and actively change’ their methodology from report to report
making comparison from year to year difficult. For example, only two indicators remain un-
changed between the fifth and sixth editions [30]. For the latest edition PWC have included
the results of a staff survey (20% response rate in each of the cities) covering part of the
Demographics and liveability area (commute and relocation attractiveness—which seems to
be a recurring theme).

2.1.3 The Economist’s Intelligence Unit’s Indices

The Economist Intelligence Unit produces a number of liveability indices on a more-or-less
annual basis. Their Liveability Survey [39] (which is again intended to be used to assign
salary loadings when relocating staff) results are calculated from a mix of ‘over 30 quali-
tative and quantitative factors across five broad categories: stability; healthcare; culture and
environment; education; and infrastructure’. There is no questionnaire for this survey, all
the results are generated by EIU experts assisted by in-city collaborators who rank the qual-
itative factors based on their experience. The quantitative factors are obtained from reliable
sources such as the World Bank. Each of the five broad categories are weighted and, inter-
estingly, Education again receives the lowest weight. The bias of this index towards English
speaking cities is well known [17].

Perhaps of more interest for current purposes is the ‘Where-to-be-born’ (or ‘Lottery of
Life’) Index [38] (previously known as the ‘Quality-of-life’ Index) which uses quantitative
data to predict qualitative survey results in a multivariate regression framework. Unfortu-
nately, this index is reported on a country-by-country basis and is thus slightly out of scope.
However, we still feel it worthwhile to discuss in a little more detail due to the methodology
employed in its generation. We base our discussion on [37] which was made available for
the original 2005 index.

The EIU uses the ‘average scores from comparable life-satisfaction surveys’ taken in
the years 1999 or 2000 from 74 countries which are then related in a multivariate regres-
sion framework over nine (almost) objective measures: Material wellbeing (PPP), Health
(life expectancy at birth), Political stability and security (EIU rating, subjective), Family life
(divorce rate), Community life (trade union membership and church attendance—binary; 0
if low; 1 if high—subjective), Climate and geography (latitude), Job security (unemploy-
ment rate), Political freedom (Freedom House [2], subjective) and Gender equality (ratio
of average male and female earnings). The outputs of the regression are used to derive the weights applied to the scores from the nine ‘objective’ measures listed above thus enabling the calculation of the requisite index. Despite listing the regression statistics the approach is incomplete: the ‘comparable life-satisfaction’ surveys are not referenced and the values of the nine ‘objective’ variables chosen to project the 1999/2000 results forward to 2005 are not available; so unfortunately, results are not reproducible. The use of linear multivariate regression (inferred from [37]) may not be appropriate: doubling PPP does not appear to be linearly related to liveability. Clearly, nonlinear techniques are required.

3 Computable Liveability

The essential idea of ‘Computable Liveability’ extends the ideas encompassed in the EIU Quality-of-Life Index by making use of modern machine learning techniques. Instead of deriving weights in a (linear) regression framework the aim is to directly estimate subjective liveability indices (whichever one we choose) directly from measured (objective) data in an unconstrained data-driven, nonlinear manner. The idea is shown schematically in Figure 1. We note that this is a non-linear regression problem where we make no constraints on the degree of non-linearity.

![Machine Learning System](image)

**Fig. 1** From objective city measurements to estimated subjective liveability

The fundamental reasoning behind the proposed approach is that subjective survey responses to liveability questions must be related to objectively measurable parameters. For example, if a city opens a new metro line that connects people together more efficiently we would anticipate a change in a subjective liveability score. The relationship is clearly...
complex and highly nonlinear; nevertheless it this function that we aim to realise through machine learning.

It must be noted that both the livability score and the objective measurements are uncertain and two of the machine learning approaches we propose will deal directly with the uncertainty in both input and output, in a principled and manageable way.

We envisage two prototypical uses.

1. Given a time series of subjective livability indices, \( L_T(t) \), for a given city and an \( n \)-dimensional vector of corresponding objectively measured quality-of-life indicators, \( Q_T(t) \) (note that the set of objectively measured quality-of-life indicators is contained in the set of quality-of-life-indicators), we seek a function \( F_T \) such that

\[
L_T(t) = F_T(Q_T(t)) : \mathbb{R}^n \mapsto \mathbb{R}
\]  

The function \( F_T \) will allow us to predict the subjective livability of the given city at some future time given the objectively measured parameters at that time.

2. Given a subjective livability index, \( L_C \), for a collection of distinct cities labelled by \( C \) and an \( m \)-dimensional vector of corresponding objectively measurable quality-of-life indicators, \( Q_C \), we seek a function \( F_C \) such that

\[
L_C = F_C(Q_C) : \mathbb{R}^m \mapsto \mathbb{R}
\]  

The function \( F_C \) will allow us to predict the subjective livability of a new city given the objectively measured parameters.

In general equality will not be met and nor will it be possible to find an explicit representation of the mapping, however, neither of these is a limitation in practice.

For both cases we have made the implicit assumption that the subjective survey responses will be directly comparable from year-to-year and city-to-city. While the former seems likely; the latter is less so. As an example there is evidence that suggests that climate has a significant influence on subjective livability [16] but will its effects on the population be similar from city-to-city?

Combining both prototypical cases into the fully general problem of time varying livability across a number of cities is possible, however, we will not discuss it here other than to point out that extending the methods discussed below to the fully general case is straightforward.

### 4 Mathematical Approaches

There are a number of mathematical techniques that can be used to estimate the functions \( F_T \) and \( F_C \). We first note that from the mathematical point of view the two prototypical cases are equivalent but the interpretation of the results is quite different. To facilitate notation we will drop the subscripts and more simply denote our desired function by \( F \) and the livability score by \( L \). Similarly we denote our vector of objectively measurable quality-of-life indicators simply by \( Q \). Thus our canonical regression problem is to find a function \( F \) such that

\[
L = F(Q) : \mathbb{R}^k \mapsto \mathbb{R}
\]  

that is valid for all QoL inputs, \( Q \).
From the supervised machine learning (ML) perspective we adopt here the problem we are solving can be cast as follows. Given a training set \( T = \{ (Q_i, L_i) : i = 1 \ldots p \} \) of objectively measurable quality-of-life indicators \( Q_i \) (each of dimensionality \( k \)) and corresponding liveability scores \( L_i \) we seek a computable liveability function \( \mathcal{F}(Q) \) that approximates the relation between the training set points as closely as possible and can be used to compute liveability scores for previously unseen inputs. Just how well it does this is known as the generalization capability of the particular algorithm used to find \( \mathcal{F} \). The greater the generalization capability the ‘closer’ \( \mathcal{F} \) will be to \( F \). We note that for prototypical case 1 the cardinality of the training set would typically be less than 20 and it may be up to 100 or so for case 2.

Ideally we would like to obtain the probability density associated with a predicted estimate of \( L \). Using the notation introduced above, we seek to learn a liveability function \( \mathcal{F} \) that will be the best estimator for \( L \) given the training set \( T \). In symbols

\[
\mathcal{F} \sim p(L|T)
\]

where \( p(L|T) \) denotes the conditional probability density estimate for \( L \) given the training input \( T \). We note that this is the formal statement for all regression problems, bearing in mind that it in many cases \( p \) is set identically to unity whence the estimate for \( L \) for arbitrary input \( Q \) can be taken as the conditional expectation (mean),

\[
\mathcal{F}(Q) = E(L|Q)
\]

in an implied probabilistic regression formalism.

In the interests of space, readability and the many excellent references (of which we cite examples in the text) we leave the detailed mathematics to the skilled reader. In particular, we almost completely avoid any discussion of overfitting (where the learned regression model follows the noise in the input at the expense of signal) and its possible solutions. The reader is directed to [15] for a brief discussion of the issues and to [26] for a tutorial on one approach to circumventing the problem.

As noted above it is more or less self-evident that the function \( F \) will be nonlinear. The first two methods we describe handle the nonlinearity directly whilst the latter two rely on the so called ‘kernel trick’, which is described below.

The usefulness of the kernel trick stems from the observation that the fundamental mathematical operations required for a large range of regression (and pattern recognition) methods can be cast as scalar products. Hence if we can find non-linear transformations that preserve the simplicity of the scalar product operation then all the linear theory can still be applied. Indeed such functions can be found and one particularly important class are termed Mercer kernels (they are kernels in the integral equation sense and the result called upon, known as Mercer’s Theorem [25]—not the same Mercer as above—, simply states that under conditions of kernel positivity it is possible to expand a self-adjoint (Hermitian) kernel in a uniformly convergent manner in terms of its eigenfunctions (eigenvectors in Hilbert Space)).

Given a nonlinear transformation function, \( \Phi(x) : \mathbb{R}^p \mapsto \mathbb{R}^q \), (in some cases the dimension, \( q \), of the transform space may be infinite and we leave the implications to the mathematically skilled reader) we seek a Mercer kernel \( k(x,y) \) such that the scalar product in the transformed space is given by:

\[
\Phi(x) \cdot \Phi(y) = k(x,y) : \mathbb{R}^q \times \mathbb{R}^q \mapsto \mathbb{R}.
\]
Indeed, it is simpler to ask what functions satisfy Mercer’s Theorem, obtain some interesting kernels and ignore the \( \Phi \)'s. Some interesting kernels are:

**Polynomial of order** \( d \)

\[
k(x, y) = (x \cdot y)^d
\]  

**(8)**

**Radial basis functions**

\[
k(x, y) = \exp\left(-||x - y||^2 / c\right)
\]  

**(9)**

**Neural networks**

\[
k(x, y) = \tanh(\kappa(x \cdot y) + \Theta)
\]  

**(10)**

With the machinery of the kernel trick, we can press onwards and discuss some ways of finding, at least implicitly, approximations to the function \( F \).

It is also worth noting at this point that implementing the methods described below can be quite challenging. However, there are many open-source and commercial offerings and we will suggest references as appropriate.

### 4.1 Random regression forests

Random regression forests [11] are one of a number of modern ML techniques that belong to the general group known as *ensemble* learners, and they are amongst the most flexible and efficient of all ML techniques. The output of a regression forest can be a simple point estimate or a probability density [13]. The latter is particularly useful because it will allow us to decide whether two or more of our computable liveability scores are statistically distinct.

The idea for regression forests stems not surprisingly, from the regression tree (which is the continuous-variable generalization of the decision tree). An example tree is shown in Figure 2. Note that this is a binary tree, each node has a maximum of two connections to nodes at a lower level in the tree.

![An example of a regression tree. The path traversed by a particular input is shown by the thicker line.](image)

**Fig. 2** An example of a regression tree. The path traversed by a particular input is shown by the thicker line.

We discuss regression trees before moving on to forests. The aim of training a regression tree is to learn a mapping from inputs to outputs using the training set. There are a number of parameters that must be learnt: the depth of the tree until an output (leaf) node is reached and the split at each node, i.e., should the connection to the next level go to the left or right child node. Each node \( j \) is assigned a so called weak learning function, \( W_j \), that is fast to
evaluate and gives a reasonable approximation to the split at that node. Typically, each weak learner will have a number of parameters that may be tuned to optimize its performance. Training automates this tuning. One choice for the weak learning function is the mean-squared function [11] and another is the probabilistic linear estimator [13], with the latter producing a fully probabilistic estimate of the output value.

Tree depth can be controlled in a number of ways: fixing the tree depth, limiting the number of nodes or by setting an output accuracy criterion. Now we are ready to enter the forest. A regression forest, an example of which is shown in Figure 3 is simply a collection of regression trees. The estimated function output, \( p(L|Q) \), is the average value of the outputs from the individual trees

\[
p(L|Q) = \frac{1}{N} \sum_{i=1}^{N} p_{\tau_i}(L|Q).
\]

(11)

Moreover, each tree can be trained independently. Which brings us to why ‘Random’ was the first word in this section: every tree in the forest is randomly different from every other tree. The randomness of the trees is introduced in the training phase and two of the most common methods are randomly selecting a subset of the training set for each tree [11] and the random subspace method due to Ho [19] which amounts to only tuning a random subset of the parameters associated with a weak learner at a particular node in a tree.

We note that in the extension to forests we have introduced an additional dilemma. How many trees should there be in a forest? See [27] for a discussion.

Open source software implementations of random forest regression are available for the R statistical software package [23, 31] and for Matlab [3]. C++ and C# implementations are available from [10].

4.2 Multi-layer perceptron regression

The multi-layer perceptron (MLP) is the simplest type of feed-forward artificial neural network. An MLP is typically represented as directed graph linking an input layer of processing units (‘nodes’) to an output layer via at least one intermediate (‘hidden’) layer: see Figure 4. A node in a lower numbered layer is connected to all nodes in the next higher numbered layer. Like all the methods discussed here an MLP can be used for both classification and
regression. In Figure 4 we have configured the MLP for regression; there are \( P \) real-valued inputs and one real-valued output. For the case in hand the output will be the computable liveability \( L \) and the input nodes will receive the measurable QoL indicators of which there are \( k \). The learned MLP will be an implicit embodiment of \( \mathcal{F} : \mathbb{R}^k \rightarrow \mathbb{R} \) which is hopefully a close approximation to the computable liveability function, \( F \).

As may be seen in Figure 4 the arcs of the graph are assigned weights that control the strength of excitation from nodes in one layer to those in the next. The output of the \( i \)-th node in the \( T \)-th layer, \( y^T_i \), is given by

\[
y^T_i = A \left( \sum_{j=1}^{N} w^{(T-1)}_{ij} y^{(T-1)}_j \right)
\]

where \( N \) is the number of nodes in the \((T - 1)\)-th layer; \( w^{(T-1)}_{ij} \) is the weight applied to the output, \( y^{(T-1)}_j \), of the \( j \)-th node in the \((T - 1)\)-th layer and \( A \) denotes an activation function which historically has been chosen to mimic the activation potential of biological neurons. Typical choices for \( A \) include the hyperbolic tangent and the sigmoid function. It has been shown [14, 22] that a three-layer MLP can approximate continuous functions on compact subsets of \( \mathbb{R}^n \) arbitrarily closely under fairly weak assumptions on \( A \). This is good news for
the case in hand because the measurable QoL indicators may be defined on closed intervals of \( \mathbb{R} \). Unfortunately, there is no general rule for choosing the number of nodes in the hidden layer (of which there is now only one), so one must simply use trial and error.

The learning process for an MLP finds the arc-weights based on the training set. The best known of the training algorithms is the so-called ‘back-propagation’ algorithm [34] which is a two stage supervised learning process that first runs a training example through the network to compute the activation potentials and then uses the computed activation potentials to generate the difference between the actual training output and the network-computed output. The difference is then used to update the weights using gradient descent or similar.

One accepted way of improving performance of an MLP is **pruning**. An MLP is pruned by deleting those nodes in the hidden layer which have very small weights into the output node. Of course, some loss of precision may result.

It should be noted that it is possible to introduce a kernel mapping into the MLP [33] with the aim of enhancing the generalization capability (not to handle nonlinearity which is intrinsically catered for). Unfortunately, it is not yet clear just which problems will benefit. However, it may be worthwhile exploring this more complicated version of the MLP.

Open source implementations of MLP may be found on the Web: see [18] or [5] for example.

### 4.3 Support vector regression

Support vector regression (SVR) is the natural extension of the support vector machine (SVM) [40] from binary classification to the approximation of continuous functions. We discuss two variants \( \varepsilon \)-SVR [36] and \( \nu \)-SVR [35].

The goal of \( \varepsilon \)-SVR is to find the function \( \mathcal{F}(Q) \) that has at most \( \varepsilon \) deviation from the actual function values at all the \( Q_i \) in the training set. As promised we skip almost all of the mathematics. Choosing \( \varepsilon \in [0, \infty) \) is tricky given that in many cases of regression we simply seek the closest fit, and do not want to limit the accuracy of approximation beforehand. The \( \nu \)-SVR formulation sidesteps this issue by automatically minimising \( \varepsilon \) based on the given training data through the parameter \( \nu \in [0, 1] \) (which effectively controls the number of support vectors—see below). We note that uncertainty (noise) in the inputs is handled by the choice of \( \varepsilon \) (or \( \nu \)) but the resulting output approximations do not provide easy access to their error bounds.

All SVM-type problems may be reduced to solving a constrained quadratic optimization problem (QOP) for which a unique solution always exists. It also transpires that the solution of the QOP can be written completely in terms of the training vectors \( Q_i \) and their scalar products [36]. Not all of the training vectors survive the solution process, but those that do are called **support vectors**.

The observation that the only mathematical operation involving the training vectors is the scalar product is of course the entry point for non-linear regression, via the kernel trick.

Well known and relatively ease to use implementations are available in SVM\textsuperscript{light} [21] and LIBSVM [12]. Both are available for download on the web.

### 4.4 Gaussian processes for regression

A Gaussian process (GP) [32] may be thought of as the generalisation of a multivariate Gaussian distribution to a function space (just as a Hilbert space is a functional generaliza-
tion of a finite dimensional vector space). In the finite case a multivariate Gaussian is fully described by a mean vector and a covariance matrix whereas in the function space generalisation a Gaussian process is fully described by a mean function and a covariance function. Formally a Gaussian process is a collection of random variables, any finite number of which have a multivariate Gaussian distribution. But what does this have to do with estimating liveability?

We first make the observation that the collection of liveability scores in our training set when arranged as a vector, \( \mathbf{L} = (L_1, ..., L_p) \), can be viewed as a single sample point drawn from a multivariate Gaussian distribution and so can be matched with a Gaussian process. For simplicity it is often assumed that the mean of this GP is zero (we can approximate this by subtracting the sample mean of the liveability scores). We further assume that the covariance function, \( k(Q, Q') \), describing the GP is ‘squared exponential’

\[
\text{cov}(F(Q), F(Q')) = k(Q, Q') = \exp\left(-\frac{1}{2}\|Q - Q'\|^2\right)
\]

We note that this has the same form as the radial basis function kernel, Equation 9 above, which may be shown to correspond to a linear regression model with an infinite number of basis functions. The choice of symbols, \( k(Q, Q') \) is hardly coincidental—nonlinearity is being handled by the kernel trick. Of course there are other possible choices for the covariance function (i.e., kernel) and we leave these at the discretion of the reader.

We write

\[
L(Q) \equiv F(Q) = GP(0, k(Q, Q))
\]

for the defining GP, where we have explicitly entered the zero mean. Now we are in good shape to answer the question posed above. To reiterate, we seek the liveability score, \( L^* \), corresponding to some new input \( Q^* \) (corresponding to a new city or a new time depending on the problem we are solving).

First we compute the covariance function between all the training examples

\[
K = \begin{pmatrix}
k(Q_1, Q_1) & k(Q_1, Q_2) & \cdots & k(Q_1, Q_p) \\
k(Q_2, Q_1) & k(Q_2, Q_2) & \cdots & k(Q_2, Q_p) \\
\vdots & \vdots & \ddots & \vdots \\
k(Q_p, Q_1) & k(Q_p, Q_2) & \cdots & k(Q_p, Q_p)
\end{pmatrix}
\]

and for our new input

\[
K_* = (k(Q_*, Q_1)k(Q_*, Q_2) \cdots k(Q_*, Q_p)) \quad K_{**} = k(Q_*, Q_*)
\]

We noted above that the liveability scores may be treated as random variables drawn from some Gaussian distribution (as is conventional we use the symbol \( \mathcal{N} \) to denote a Gaussian or Normal distribution) to which we assign zero mean and covariance \( K \) given in Equation 15. So

\[
\begin{pmatrix}
\mathbf{L} \\
\mathbf{L}_*
\end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K & K_*^T \\ K_* & K_{**}\end{pmatrix}\right)
\]

where superscript \( T \) is used to denote matrix transposition. Now what we really have to estimate is the conditional probability \( P(L_*|L) \): how likely is a particular estimate \( L_* \) given the previous (training) data. It can be shown that this conditional probability is Gaussian

\[
L_*|L \sim \mathcal{N}(K, K^{-1}L_* - K_*K^{-1}K_*^T),
\]
and so the best estimate for $L_*$ is the mean of this distribution

$$L_* = K_*K^{-1}L.$$  \hfill (19)

Moreover we also obtain an estimate for the variance of $L_*$

$$\text{var}(L_*) = K_{**} - K_*K^{-1}K_*^T.$$  \hfill (20)

This is indeed an elegant result. However, we have assumed that all the data we are working with is precise, which will never be the case in the real world. It must also be observed that the covariance function will, in general, have some free parameters (which are termed hyperparameters) and these may be estimated from the training data as well. We leave it up to the reader to consult the reference above for the details and suggest [4] for software and additional references.

5 Research Plan

When we started thinking about the subject of this paper we were struck by two fundamental challenges. The first was which liveability index to choose and the second was what set of objectively measurable QoL indicators to match the index with. The first issue is still open but fortunately the International Standards Organisation has gone a long way to answering the second.

The recently published (and previously cited) ISO Standard 37120 *Sustainable development of communities—Indicators for city services and quality of life* defines forty-six objectively measurable QoL indicators across sixteen areas: Economy, Education, Energy, Environment, Finance, Fire and emergency response; Governance, Health, Safety, Shelter, Solid Waste, Telecommunication and innovation; Transportation, Urban planning, Wastewater, and Water and sanitation. The reader is referred to the document for the QOL indicators defined for each area. These QoL indicators seem to be objectively measurable and unbiased.

We adopt a two stage research plan which we describe below.

5.1 Stage 1

This is a proof of concept Stage and its outcome will be used to justify the need (or otherwise) for Stage 2. We address the two prototypical cases detailed above. In both cases we will use the *Mercer* index as the subjective liveability index that we are attempting to approximate. The reason for this is that out of the three indices we examined it is the only one which is purely subjective and appears to be more or less consistent over the past 10 years. A quick search of the web will give *Mercer* scores back to 2004.

For prototypical case 1, we select a number of cities that are represented in the last ten ($n$) or more *Mercer* rankings and extract their score for each year. For each of these cities find the forty-six ISO measurable QoL indicators for each of the $n$ years. Discard cities for which the data sets are incomplete or find a reduced set of QoL indicators which span a greater number of cities or time-period. Train each of the models on $(n - 1)$ years of data reserving one year for comparison. Select the best model for each of the chosen cities by determining which model produces the closest prediction to the *Mercer* score for the reserved year.
For prototypical case 2, for all years we select the scores for all the representative cities—say year \(i\) has \(M_i\) cities listed—and for each of these \(M_i\) cities find the forty-six ISO measurable QoL indicators discarding cities or reducing the number of indicators as stated above (say \(M_i(k)\) cities survive in each year). For each year, \(i\), train each of the models on \(M_i(k) - 1\) cities; reserving one for comparison. Compare the predicted liveability score for the reserved city with the reported Mercer score. The choice of the city to leave out could be used as another test parameter.

We also need to estimate the uncertainty in liveability score and measured data as best as possible.

5.2 Stage 2

Assuming Stage 1 is successful. Stage 2 would entail a reassessment of the subjective survey data collection process. In particular a stated representative sample size would need to be selected for each city. In a similar vein online survey collection will clearly produce a biased sample especially in less developed countries. The selection of a suitable group of test cities and the development of a culturally unbiased survey is also critical for success and clearly requires strong involvement from the social science community. The survey results would be used to generate a new Liveability Score.

In parallel with survey development it will be necessary to collect and validate the forty-six ISO QoL indicators across the selected test cities.

Where there is insufficient data to do a temporal study for many years the focus would be on predicting survey results for unsurveyed cities thus removing the cost of having to carry out a survey.

6 Conclusion

In this paper we have presented a brief discussion of three of the more prominent liveability indices and discussed their differing characteristics. We have introduced the concept of Computable Liveability that aims to link subjective liveability survey scores with objectively measurable Quality of Life Indicators and presented a number of ways in which computable liveability might be estimated. We have also shown the steps needed for a research plan.

Using Computable Liveability as a policy assessment tool should enable city authorities to have an additional input into understanding the effect on the citizens of particular interventions: for example, will adding a new tram line at the expense of parkland increase or decrease the liveability as perceived by the citizens?

It is worth mentioning that the idea of estimating subjective results from objective data may apply equally well to the individual question scores in a liveability survey, and also to other subjective responses that are driven by measurable, objective inputs.

Acknowledgements NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program. NICTA is also funded and supported by the Australian Capital Territory, the New South Wales, Queensland and Victorian Governments, the Australian National University, the University of New South Wales, the University of Melbourne, the University of Queensland, the University of Sydney, Griffith University, Queensland University of Technology, Monash University and other university partners.
References

5. OpenNN. http://www.intelnics.com/opennn
Advancing sustainability indicators through text mining: a feasibility demonstration

Samuel J. Rivera\textsuperscript{a} (srivera2@illinois.edu), Barbara S. Minsker\textsuperscript{a} (minsker@illinois.edu), Daniel B. Work\textsuperscript{a} (dbwork@illinois.edu), Dan Roth\textsuperscript{b} (danr@illinois.edu)

\textsuperscript{a}Department of Civil and Environmental Engineering, University of Illinois at Urbana-Champaign, \textsuperscript{b}Department of Computer Science, University of Illinois at Urbana-Champaign

ABSTRACT

Sustainability indicators are metrics that are used to assess and track sustainable development, such as the number of people living in poverty or the conservation status of endangered species. Defining sustainability indicators is challenging because the studies are often expensive and time consuming, the resulting indicators are difficult to track, and they usually have limited social input and acceptance, which is a critical element of the social component of sustainability. The central premise of this work is to explore the feasibility of identifying, tracking and reporting sustainability indicators by analyzing unstructured digital news articles with text mining methods. Using San Mateo County, California, as a case study, a non-mutually exclusive supervised classification algorithm with natural language processing techniques is applied to analyze sustainability content in news articles and compare the results with annual sustainability indicator reports created by Sustainable San Mateo County (SSMC) using traditional methods. The results showed that the text mining approach could identify all of the indicators highlighted as important in the SSMC reports and the method has potential for identifying region-specific sustainability indicators, as well as providing insights on the underlying causes of sustainability problems. Some chronic problems that are considered less newsworthy proved more difficult to track with news articles and may be better monitored using other types of online media, such as blogs and reports. Such use of online media can improve the incorporation of society’s values in the process of selecting and tracking the indicators, a component that had been missing in previous approaches.

KEYWORDS: sustainability indicators, text mining, informatics, knowledge discovery
1. INTRODUCTION

Sustainability indicators are metrics that track the current state and evolution of complex systems (Hammond et al., 1995; IISD, 2000), such as the number of people living in poverty or the health of endangered species. To be comprehensive, these indicators must address the political, economic, social, and environmental components of communities, and must be understood by all members of society (Innes et al., 2000; Dahl, 2012). Additionally, indicators should be aligned with the values and concerns of the target audience (Dahl, 2012) and be created at multiple scales of governance (i.e., global, national, regional and city scales) (Bossel, 1999; Innes et al., 2000; Gahin et al., 2003; Dahl, 2012).

To date, 895 initiatives exist worldwide to develop sustainability indicators ranging in scales from cities to global projects (IISD, 2013). However, not all of the methodologies and guidelines for developing and implementing sustainability indicators have been effective (Gahin et al., 2003; Krank et al., 2011). Most indicator projects have often failed to achieve large-scale participation, and consequently may not be representative of the community’s true values and concerns (Innes et al., 2000; Gahin et al., 2003; Adinyira et al., 2007; Yli-Viikari, 2009; Scerri et al., 2010; Krank et al., 2011; Dahl, 2012; Moldan et al., 2012). Furthermore, the collection and analysis of this input is often extremely time consuming and resource intensive, which ultimately leads to large latencies in the reported data (Innes et al., 2000; Gahin et al., 2003; Moldan et al., 2012).

The objective of this work is to begin addressing these limitations through the development of a new method to design and track sustainability measures using digital news media and recent progress in text mining. More specifically, this project addresses the question of whether the news media contains relevant information that can enable fast identification, tracking, and reporting of sustainability indicators for a region. The hypothesis to be tested is that the unstructured data of news media can provide insight into sustainability problems within the cultural and contextual characteristics of a community, thereby also addressing the social component that has been underdeveloped in previous approaches.
News media represents a “mediated public sphere” (Holt et al., 2012) with the potential to influence people’s mindsets and create a feeling of worldwide connectedness by changing the public’s level of awareness and attention to a specific issue (Szerszyski et al., 2000; Holt et al., 2012). Moreover, it has been argued that there is a causal relationship between thematic priorities of the media and the relevance of social problems in the population (Roggers, et al., 1993). Furthermore, news articles contain far more than just factual details; they provide insights into the cultural context upon which they are written, a spatial and temporal component to the facts, and a window for forecasting many social behaviors using text mining techniques (Thøgersen, 2006; Tang et al., 2009; Leetaru, 2011; Michel, et al., 2011).

Text mining tools have often been used to process large volumes of unstructured digital news to identify actionable information and extract trends. Recent studies suggest that analysis of text archives can generate new knowledge about the functioning of society (Leetaru, 2011). Moreover, text mining tools can detect the tone of news articles, which enables applications such as forecasting social behaviors ranging from ticket movie sales to stock market trends (Mishne et al., 2006; Tang et al., 2009; Leetaru, 2011; Michel, et al., 2011). This suggests the use of text mining techniques to quantify and assess the social components of sustainability could hold promise.

1.1 CONTRIBUTIONS

Previous sustainability studies using text mining have focused on tracking general trends in different sustainability and climate change topics at the national or global scale (Barkemeyer, et al., 2009; Scharl et al., 2013). This study focuses on demonstrating a faster method for identifying, tracking, and reporting of sustainability indicators specific to a region using news articles that can better incorporate society’s values. Furthermore, the approaches taken in this study provide links between observed indicator trends and their underlying causes; previous indicator methods focus primarily on tracking the issues (Barkemeyer, et al., 2009; Scharl et al., 2013).

Additionally, this study develops a new methodology that combines a suite of text mining methods to more accurately classify sustainability articles given the limited training set of
regional news articles and their non-mutually exclusive topic areas (e.g., an article on the effects of pesticides on water quality would be equally relevant to the water quality and pesticide indicators). The methodology is applied in San Mateo County, California (CA), to demonstrate feasibility of the approach. Future work is then recommended to extend the methodology to other types of data available on the Web (e.g., social media data and blogs). Finally, it is important to mention that this paper only provides a brief summary of the project, as further details can be found in our peer-reviewed article (Rivera et al., 2013).

2. METHODOLOGY

The tracking and extracting of information from sustainability related news articles is accomplished by integrating different classification approaches and natural language processing (NLP) techniques. The methodology implemented in this project had three main components (Fig. 1): pre-processing of the unstructured textual news data, classification of the documents under predetermined labels (i.e., sustainability indicators), and NLP information retrieval.

In the pre-processing step the textual data from a set of documents is represented by a word-document matrix. The matrix is constructed by assigning an importance metric to each word in the article (e.g. frequency of occurrence of the word in the document). These metrics are then stored in the document’s row of the matrix. In this study the binary representation of the occurrence of a word and the term frequency-inverse document frequency (TF-IDF) were used as metrics of importance. This word-document matrix representation allows the calculation of similarity between the new document and a pre-labeled set, which is used for classification of the document under the most related sustainability indicators.

The classification of news articles under the sustainability indicators using typical supervised classification algorithms posed new challenges due to the non-mutually exclusive characteristics of the sustainability news articles, which often span multiple sustainability indicators, and the manual human labor associated with the task of pre-labeling a large training set. To address these challenges the classification of news articles into different sustainability indicators was done using two supervised classification algorithms, K-Nearest Neighbors (KNN) and Dataless Classification (Chang et al., 2008), and a sustainability indicators hierarchical
category tree. The hierarchical category tree structures was used to reduce the possibility of misclassification due to similarity between words in documents that are not in the same category (e.g., a news article talking about wind power versus one taking about a storm) (Sun, et al., 2001, Silla Jr., et al., 2011), thus reducing the non-mutuality of sustainability-related news articles. The classification of documents within this hierarchy is composed of three main steps: (1) transformation of the data using a generalized discriminant analysis (GDA) (Li et al., 2008) to reduce dimensionality of the classification problem, (2) classification of documents as related or non-related to the sustainability indicators using K-Nearest Neighbors (KNN) and (3) classification of the documents under the parent and child nodes of the category tree using KNN, Dataless classification, and a majority voting rule. Given the large set of sustainability indicators, the approach uses KNN to suggest candidate class labels to the Dataless classification algorithm (Chang et al., 2008), which takes only two labels as input and uses Wikipedia as its training set to classify the news articles. The majority-voting rule then combines the KNN and Dataless results to determine a final classification of each article. This novel combination of these techniques aims at reducing the time required for the task of pre-labeling news articles and allowing the use of a small training set in the classification model.

Finally, NLP techniques are used to assign a geographical location and identify co-occurring topics in the set of documents. Two main techniques are used to complete this task: (1) the creation of frequent concept sets and association rules that identify and summarize recurrent topics across all news articles classified under the same sustainability indicator, and (2) Part-of-Speech (POS) tagging and a gazetteer for geo-referencing of news articles, identifying areas with the highest interest in a particular sustainability indicator. This information can provide additional insights that help explain the current state of the sustainability indicators.
3. CASE STUDY APPLICATION AND DISCUSSION

To illustrate the feasibility of using news articles to identify and track sustainability indicators, a case study was examined in the region of San Mateo County, CA. This location was selected because of the availability of a digitized local newspaper archive and its long-standing sustainability indicator program, which is led by the Sustainable San Mateo County (SSMC) non-profit public benefit corporation. The SSMC has generated sustainability indicator reports annually since 1998 where the current states of 49 sustainability indicators are presented to the community. The current states of these indicators are defined by a series of individually tracked sub-indicators that have specific data associated with each of them.

The feasibility analysis conducted in this case study included 22 sustainability indicators and 36 sub-indicators. The majority of the selected indicators were extracted from the SSMC sustainability reports based on the expected likelihood of their appearance in the news articles. The rest of the indicators were selected from other sustainability reports of international agencies.
(e.g., United Nations), and regional reports from different geographic areas. These additional indicators represented sustainability problems often seen in similar coastal regions, and were included in the analysis as a validation to the proposed methodology. Lastly, the analysis was conducted using news articles from the San Mateo County Times newspaper. The San Mateo County Times is among the newspapers with the highest circulation in the region, and the only one with an accessible digitized archive. The results of the case study include a performance evaluation of the classification approach and a set of illustrative results that demonstrate the feasibility of using news articles to identify and track sustainability indicators. To validate these illustrative results the SSMC reports were used as a representation of the experts’ opinions.

The classification accuracy, recall and precision were used to assess the performance of three different supervised classification methods (i.e., K-nearest Neighbors, Support vector machine and Naive Bayes) at the root (binary classification scheme) and parent nodes (multi-class classification scheme). The results demonstrated that KNN outperformed the other classification methods; therefore KNN was selected as the preferred classification algorithm for these levels of the hierarchical tree. At the root node (classifying sustainability vs. non-sustainability articles), the classification approach achieved 88% accuracy with 90% recall and 85% precision for sustainability-related articles. However, at lower levels of the multi-label tree, the overlap of words becomes higher, producing less distinct classes and a harder classification problem. Thus, at the parent and child node level, the method achieved 88% and 71% accuracy, respectively. Nevertheless, the performance is satisfactory given that the percentage agreement between humans performing the same task was only 86%. Moreover, the classification accuracy at lower levels of the hierarchical tree proved to be sufficient to allow extraction of meaningful information from news articles.

To demonstrate the method’s feasibility of identifying and tracking the indicators using news articles, each of the 2007 and 2009 articles from the San Mateo County Times were classified under one of the sustainability indicators, and the volume of news articles was used as a measure of the community’s perception of importance of each indicator. This measure was based on the use of volume of media coverage as a traditional measure of the relative importance of an issue over time in previous studies (Mueller, 1973; Benton and Frazier, 1976; Naisbitt, 1976; Beniger, 1978; Rogers et al., 1993; Dearing and Rogers 1996; Carvalho, 2005; Holt et al.,
The indicators were also tracked by calculating the monthly volume of news articles using their date of publication. Additionally, information retrieval techniques were used to analyze months with unusually high numbers of articles to identify events that led to such high volumes. Finally, the ability to locate the sustainability problems within the region of San Mateo County was evaluated by using the POS techniques and the GeoNames database (www.geonames.org/) to geo-reference news articles to a specific area within the county.

In summary, the results of this work indicate significant promise for text mining to address the timely and resource-intensive nature of identifying, tracking, and reporting the indicators by providing regular updates (e.g., daily, weekly, or monthly) on: (1) the relative importance of different sustainability indicators to a region, (2) the current state of these indicators and (3) underlying events affecting them. The results demonstrated the capabilities of the method for identifying sustainability issues specific to a region by providing insights on the relative importance given to different sustainability indicators by the community, an issue that had been overlooked in the past and which is necessary for the creation of value-based indicators. Additionally, the analysis of data to provide an explanation to the observed changes is likely one of the most human intensive and time consuming tasks associated with indicator reporting. The method was able to provide the same explanations for the behavior of the sustainability indicators in near real time by suggesting associated keywords and correlation between many of the indicators. This additional information allowed the identification of the events causing shifts in discussions of these sustainability issues (e.g., the Cosco Busan oil spill and the decrease in the air quality due to particulate matter due to wood burning in the winter months). All of these results correlated well with the information provided in the SSMC sustainability reports, thus validating the capabilities of the method for tracking and reporting the state of the sustainability indicators.

However, the results of the study were limited by sparse data that prevented the identification of some of the more chronic problems in the region (e.g. loss of endangered species, crop loss due to invasive species, and carbon emission due to vehicles). These types of sustainability issues are less newsworthy, limiting the approach to acute events that are more likely to be frequently reported. Other online news sources (e.g., blogs associated with citizens groups that are concerned about a particular chronic topic) may be more fruitful for tracking
chronic problems. Nonetheless, the low frequency of news articles related to chronic sustainability issues suggests that society may place a lower value on chronic problems.

4. CONCLUSIONS

This study is a first attempt at utilizing news media to more easily create effective sustainability indicators. Using a non-mutually-exclusive document classification algorithm and the incorporation of different information retrieval techniques, the analysis demonstrated that mining the growing amount of digitized news media can provide useful information for identifying, tracking, and reporting sustainability indicators. Using San Mateo County as a case study, the results suggest that the approach has potential for reducing time and resources dedicated to the identification and tracking of sustainability problems specific to a region. Furthermore, the use of news media allowed incorporation of society’s values in the process of selecting the indicators, a component that had been underdeveloped in previous approaches. Additionally, the method was also able to detect significant changes in the sustainability indicators and provide an explanation for such behavior in real time. All of these capabilities provide decision-makers with faster access to reliable information on the sustainability indicators at lower cost, enabling better planning aligned with the values and concerns of the community.

The limitation of the method proved to be the identification and geo-referencing of some of the most chronic problems in the region. This limitation can be attributed to the nature of the data given that not all topics are newsworthy all the time. Further research is needed to explore whether the use of additional social media sources (e.g., environmental blogs or Facebook groups) could provide more information on these chronic problems, or whether input from professionals and data analysis are still needed for these problems. Furthermore, in order to more thoroughly assess the method’s utility, its performance should be expanded to include different data sources, longer time series, and different regions across the nation.

Data ownership and licensing of news media articles were a major challenge in this study that prevented such an analysis at this time. Nonetheless, current trends point towards an easier dissemination of news data in the future, and thus a larger potential to further investigate the approach.
ACKNOWLEDGEMENTS

We acknowledge the financial support of the College of Engineering Support for Under-Represented Groups in Engineering (SURGE) Fellowship program and the Campus Research Board Program at the University of Illinois at Urbana Champaign.

SOFTWARE AVAILABILITY

Free download with installation manual and supporting material can be found at the GitHub account of the Environmental Informatics and Systems Analysis group at the University of Illinois at Urbana-Champaign (https://github.com/EISALab). The source codes is licensed under The University of Illinois/ National Supercomputer Application Center (NCSA) Open Source License (http://opensource.org/licenses/NCSA).

REFERENCES


Benton, M., & Frazier, P. (1976). The agenda setting function of the mass media at three levels of 'information holding'. Communication Research, 3(3), 261-274


Combining Physical and Participatory Sensing in Urban Mobility Networks

Xiao-Feng Xie $^{a,c}$, Zun-Jing Wang $^{b,c}$

$^a$ The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA
$^b$ Department of Physics, Carnegie Mellon University, Pittsburgh, PA
$^c$ WIOMAX LLC, Pittsburgh, PA

xie@wiomax.com, wang@wiomax.com

Abstract

It is very important to understand human mobility and activity patterns in urban environments. In smart traffic control systems, abundant traffic flow data could be collected over time by physical sensing. However, each controlled region only covers a small area, and there is no user information in the data. The rapid rise of location-based services provides another opportunity to achieve the information of human mobility, in the form of participatory sensing, where users can share their digital footprints (i.e., checkins) at different geo-locations (i.e., venues) with timestamps. These checkins provide a broad citywide coverage, but the instant number of checkins in urban areas is still limited. In this study, we focus on exploring the potentials of combining physical and participatory sensing data in urban mobility networks, based on 3.4 million checkins collected in the Pittsburgh metropolitan area, and 125 million vehicle records collected in a sub-area controlled by the SURTRAC adaptive traffic control system. First we display the spacial and temporal characteristics of the sensing data. Next, we perform user checkin statistics to reveal the distribution of user behaviors, and study entropy and regularity by cluster analysis to show a strong time-dependent predictability in human mobility patterns. Finally, we illustrate some potential usages of the sensing data in urban mobility applications, e.g., finding reasons in anomaly traffic detection, disclosing nontrivial traffic-related information in topic-specific checkins, and providing traffic origin and destination patterns based on transitions between user checkins. This work provides some essential supports for improving urban mobility.

Index terms — Participatory Sensing, Human Mobility, Urban Mobility, Traffic
1 Introduction

Understanding human mobility and activity patterns in urban environments is a significant and fundamental issue from various perspectives (Brockmann, Hufnagel, & Geisel, 2006; Simini, González, Maritan, & Barabási, 2012; Song, Koren, Wang, & Barabási, 2010; Gonzalez, Hidalgo, & Barabasi, 2008; Han, Hao, Wang, & Zhou, 2011), e.g., understanding regional socio-economics, improving traffic planning, providing local-based services, and promoting sustainable urban mobility. Traditionally, relevant information is however rarely obtained due to difficulties and costs in tracking the time-resolved locations of individuals over time.

In recent years, an increasing attention has been placed on introducing smart traffic control systems into urban road networks (Papageorgiou, Diakaki, Dinopoulou, Kotsialos, & Wang, 2003; Xie, Smith, Lu, & Barlow, 2012). The primary objectives of such systems are to reduce travel time, resolve traffic congestion, and reduce vehicle emissions. Recent work in real-time, decentralized, schedule-driven control of traffic signals has demonstrated the strong potential of real-time adaptive signal control in urban environments (Xie, Smith, Lu, & Barlow, 2012; Xie, Smith, & Barlow, 2012). The system, called SURTRAC (Scalable URban TRAffic Control), achieved improvements of over 26% reductions in travel times, over 40% reductions in idle time, and a projected reduction in emissions of over 21%, in an initial urban deployment (Xie, Smith, & Barlow, 2014). To facilitate effective real-time control, vehicle flows are monitored by different physical sensors, e.g., induction loops and video detectors, and pedestrian flows might be detected and inferred using push-buttons or other devices. Traffic flow data in fine granularity can be logged in real time, but usage of these physical sensing data is often limited to the region that is being controlled. For the broad uncontrolled regions instead, no information is available.

The rise of location-based services provides another way to achieve the information of human mobility, in the form of participatory sensing (Burke et al., 2006; Silva, Melo, Almeida, & Loureiro, 2013; Doran, Gokhale, & Konduri, 2014). With mobile devices, users can share their digital footprints at various geo-locations (i.e., venues) with timestamps through checkins, e.g., geo-enabled tweets and geo-tagged photos and videos. Using of these services grows fast worldwide, although the instant sampling rate of trajectories is still very limited, and some web-based services, e.g., Waze and Facebook, do not open their location-based data to public. Different work has been conducted to understand temporal, spatial, social patterns, and some combined patterns of human mobility (Ferrari, Rosi, Mamei, & Zambonelli, 2011; Chiang, Lin, Peng, & Yu, 2013; Wang, Pedreschi, Song, Giannotti, & Barabasi, 2011; Z. Cheng, Caverlee, Lee, & Sui, 2011; Gao, Tang, Hu, & Liu, 2013; Arase, Xie, Hara, & Nishio, 2010; Cranshaw, Schwartz, Hong, & Sadeh, 2012; Noulas, Scellato, Mascolo,
In this study, we focus on exploring the potentials of combining physical and participatory sensing data in urban mobility networks. The participatory sensing data are collected in the Pittsburgh metropolitan area with the APIs provided by the location-based services including Twitter, Foursquare, Flickr, Picasa, and Panoramio. The physical traffic flow data are collected in an area controlled by the SURTRAC adaptive traffic control system.

We first display the basic spatial and temporal characteristics of the physical and participatory sensing data. Afterward, we study human mobility patterns. User checkins are examined to disclose the distribution of user behaviors, which is a fundamental statistical properties of mobility pattern. geo-location based cluster analysis is performed to identify personal favorite places of users in the studied regions. User entropy is measured to reveal the degree of predictability of user activities. Time-dependent mobility patterns are analyzed to show the regularity of user behaviors, based on most visited places of users.

Finally we presented results of combining physical and participatory sensing data in urban mobility applications. We evaluate the attraction and limit in using sensing data on anomaly detection and reasoning for the time series of traffic flow. We examine if nontrivial information could be extracted from participatory sensing data to effectively recognize traffic congestion in temporal and spacial dimensions. We then choose two zones in the controlled region to check closely on the correlation between physical and participatory sensing patterns. For the two zones, we also investigate the origin and destination (O-D) patterns from the transitions between user checkins, which are valuable for urban mobility.

2 Data Description

We implement our study in the Pittsburgh metropolitan area. The participatory sensing data contain a list of checkins. Each checkin can be represented as a tuple $<userID, venueID, time, [comment]>$, where $userID$ is associated with a unique user, $venueID$ is associated with a venue at the geo-location of $(latitude, longitude)$ with the precision of six decimal places. Our checkin data were collected (between March and July of 2014) from the geo-APIs of some location-sharing services, including geo-enabled tweets from Twitter and geo-tagged photos from Flickr, Picasa and Panoramio. We also included existing checkin data directly crawled from Foursquare (Long, Jin, & Joshi, 2012). For studying urban mobility patterns, we only consider checkins at venues in the spacial latitude/longitude bounding box of $(40.309640, -80.135014, 40.608740, -79.676678)$, as shown in Figure 1a. For studying up-to-date patterns, we only consider the recent data within the range of dates $[1/1/2012, 7/1/2014]$. The collected data contains 3,399,376 checkins of 74,658 users at 2,198,572 venues.
The physical traffic flow data are collected in a road network which is currently controlled by an adaptive traffic control system called SURTRAC (Xie, Smith, Lu, & Barlow, 2012; Xie, Smith, & Barlow, 2012, 2014) (see Figure 1a, and for more details see Figure 1b). The system was first installed on nine intersections (A to I) in the East Liberty neighborhood since June, 2012, and then expanded to nine more intersections (J to R) in the Bakery Square neighborhood since October, 2013. The total area include five major streets, Penn Ave, Centre Ave, Highland Ave, East Liberty Blvd, and Fifth Ave, with dynamic traffic flows throughout the day. For collecting real-time flow data, detectors were deployed on each entry/exit lane at the near end of each intersection and on entry lanes at the far end of boundary intersections. For each detector, a vehicle record is generated at the time when a vehicle is detected to pass the detector. Compared to that in a checkin, there is no userID and comment information available in a vehicle record. In total, the traffic flow data contains 125,369,318 vehicle records generated at 126 stop-bar detectors by the end of 7/1/2014.

3 Basic Spacial and Temporal Characteristics

3.1 Spatial Distribution of Checkins

We first investigate the spatial distribution of all checkins in the whole region. Figure 2a is the geo-distribution of checkins. It shows a highly non-uniform dynamics of human mobility in the urban area. A few high-density regions (red colored) are shown in the heat map (see Figure 2b), one of which overlaps with the controlled region in Figure 1b.
3.2 Temporal Patterns

We are interested in the recurrent nature of human mobility over time. To investigate temporal mobility patterns, each week is segmented into $24 \times 7 = 168$ hourly bins (starting from Monday). For obtaining seasonal average results, each year is divided into four seasons (A to D), and the binned results are averaged over 13 weeks in each season. We considered all four seasons in 2013 and the first two seasons in 2014.

Figure 3a gives the seasonal average checkin patterns in the participatory sensing region. It shows that the number of checkins has increased significantly over the seasons. Figure 3b shows the average checkin frequency normalized by the total checkin size in each season. It shows that the checkin frequency has similar patterns for different seasons. The “social day” (Silva et al., 2013; Noulas et al., 2011) of Pittsburgh starts at around 4AM, and the checkin frequency peaks at around 8-9PM. It also shows high checkin activity during Sunday.

For vehicle flow in the controlled region (Figure 1b), we considered two pivotal intersections, i.e., intersection D of Centre Ave and Penn Ave in East Liberty and intersection P of Fifth Ave and Penn Ave in Bakery Square, which service for most vehicles in this road network. The seasonal average vehicle flow patterns are shown in Figures 4a and 4b, respectively. For intersection P, the vehicle flow data is only available for the two seasons in 2014. During weekdays, the traffic flow has three peaks in the morning (8AM), middle day (12PM), and afternoon (5PM). During weekends, the traffic pattern presents a smoother change that peaks at around 1PM. The flow in weekdays is heavier than that in weekends.
4 Human Mobility Patterns

In this section, we evaluate human mobility characteristics based on user activities.

4.1 User Checkin Statistics

Figure 5 presents some basic user checkin statistics in the participatory sensing region. Let $N_i$ be the number of checkins for user $i$, it turns out that the probability distribution follows a scaling law (see Figure 5a). We then consider the distribution of the radius of gyration ($r_g$) for each user. For user $i$, let $V_i^j$ be the venue of the $j$th checkin, then $r_g^i = \sqrt{\frac{1}{N_i} \sum_{j=1}^{N_i} (\text{dist}(V_i^j, V_c^i))^2}$, where the $\text{dist}$ function gives the distance between a checkin $V_i^j$ and the center of mass $V_c^i$ for the checkins of user $i$. As shown in Figure 5b, most user activities are confined to a limited neighborhood within 10 km, which is similar to the findings in the ref. (Gonzalez et al., 2008). Figures 5c and 5d show the distributions in the time and distance intervals between consecutive checkins made by users, respectively. The probability of inter-checkin times decreases with the increase in time, and interestingly, it has apparent daily and weekly patterns. The distribution of inter-checkin distances is quite similar to the distribution of $r_g$ in Figure 5b.
4.2 Finding Checkin Places

Each user tend to stay in a limited number of places, where each of places is defined to accommodate similar checkins/activities of the user in its vicinity (C. Cheng, Yang, King, & Lyu, 2012; Gao et al., 2013; Song, Qu, Blumm, & Barabási, 2010). This definition of places is able to tolerate some geo-location tracking errors. Notice that the existing tracking techniques (e.g., GPS, WiFi, and mobile tower) might have the location inaccuracy as high as several hundred meters (Jiang, Ferreira Jr, & Gonzalez, 2012; Song, Qu, et al., 2010).

For our collected data, we identify the checkin places by clustering, where each cluster $C$ defines the place of a set of checkins. We use an unsupervised clustering method, DBSCAN (Ester, Kriegel, Sander, & Xu, 1996), which is a density-based clustering algorithm that requires only two parameters, i.e., $\text{eps}$ to define a neighborhood threshold and $\text{minPts}$ to define a density threshold. By default, $\text{eps} = 250$ meters and $\text{minPts} = 2$ points were used.

4.3 User Entropy

User entropy is a fundamental quantity to capture the degree of predictability for a user (Song, Qu, et al., 2010). For user $i$, let $C_i^k$ be the $k$th cluster, and $K$ be the number of clusters. After the clustering, the temporal-uncorrelated user entropy $S_i$ is defined as $S_i =$

Figure 4: Seasonal Average Vehicle Flow Patterns for The Two Intersections in Figure 1b.
\[ -\sum_{k=1}^{K} p_i(k) \log_2 p_i(k), \] where \( p_i(k) = \frac{|C_i^k|}{\sum_{k=1}^{K} |C_i^k|} \) is the probability that cluster \( k \) was visited by user \( i \). A lower user entropy means a higher degree of predictability for visitation Patterns. Figures 6a and 6b respectively show the user entropy versus the checkin numbers \( (N) \) and the cluster numbers \( (K) \). The users in a single cluster is not shown in the log-scale figures since their user entropy are zero. As shown in Figures 6a and 6b, a larger number of checkins might lead to a lower user entropy, while a larger cluster number often leads to a higher user entropy.

### 4.4 Regularity

We study the mobility regularity of users by computing the probability of finding users in their primary and secondary most-visited places (ranked using \( p_i(k) \)) at hourly interval in a week, using DBSCAN clustering with \( eps = \{250, 1000\} \) meters (see Figure 7).

We first check the case for the primary most-visited places. The regularity is high during the night, and it peaks at around 4AM, the start of a “social day”. In weekdays, there are minima during morning (around 8AM), middle day (1PM), and afternoon (6PM) corresponding to the transitions to other places for commuting or having lunch. In weekends, there is no local minimum during morning. This pattern indicates that the primary most-visited
places should contain a large portion of home places.

For the secondary most-visited places, the regularity curve peaks at around 8AM, and reaches minima during night in weekdays, but appears being quite flat in weekends. Therefore, the secondary most-visited places are likely associated with work places.

Notice that when we change $\epsilon$ from 250m to 1000m (i.e. clustering becomes more coarsed), the increase in probability for the primary most-visited places is much more than the decrease in probability for the secondary most-visited places. This means that many users perform their other activities (i.e. except for the primary and secondary activities) near their home places (primary places). During night, the probability is nearly 90% for the primary place and 10% for the secondary place.

5 Urban Mobility Applications

In this section, we explore the usage and limit of combining physical and participatory sensing data in applications of urban mobility networks by presenting a few examples.
5.1 Anomaly Detection and Reasoning

In 2013, the bridge on South Highland Ave (location shown in 1b) was closed for replacement. This cut off the connection between Shadyside and East Liberty. Therefore a large portion of traffic (including all bus lines) on Highland Avenue was forced to pass through intersection D, which is a pivotal node that services for most vehicles in this road network.

Figure 8 shows the vehicle flow pattern in year 2013 for the right-turn movement at intersection D (see Figure 1bb, from C to D to E). The flow significantly increased between early March and late October. This is a typical change point problem for anomaly detection in time series of traffic. Using analysis techniques of statistics (Hajji, 2005) on traffic data, we can approximately detect change points, but we are not able to find the exact time and the reason leading to the events.

Analyzing participatory sensing data is helpful at this point. A search of our geo-tagged data gives five checkins with the time of the bridge reopening ceremony. For example, at 5:47PM, October 23, a user mentioned joyfully, Wow the bridge on South Highland is open. Life was rough for a while. We did not find a checkin associated with the bridge closure for our collected data, which might due to the closure event happened too long time ago and the earlier data was not kept by our location-sharing services. A search of the non-geo-tagged Tweets pinpoints the time both for closure and reopening events (for which nine and fourteen tweets have been found respectively, posted by some of the users or their first-degree friends). In further studies, fusing non-geo-tagged and geo-tagged information will be very useful for anomaly detection and reasoning, especially in case the available geo-tagged checkin information is not sufficient. In fact, many users generate both non-geo-tagged and geo-tagged information in practice. In presence of a high user regularity as shown in Section 4.4 (which is a very common case), the locations of non-geo-tagged information of the users can be inferred from their geo-tagged checkins. In addition, locations can also be estimated.
using content-derived information (Z. Cheng, Caverlee, & Lee, 2010).

5.2 Topic Related to Traffic

Many users generate checkins with the information of their current traffic conditions in travel. Tweet semantics has been used to build indicators for long-term traffic prediction (He, Shen, Divakaruni, Wynter, & Lawrence, 2013). By taking the search topic “traffic” as an example, we examine more potential usages of the participatory sensing checkins for obtaining important traffic information. Figures 9a and 9b show the spatial patterns of the checkins. A global view (see Figure 9a) indicates most checkins are located on major highways and congested neighborhoods. While a local view (see Figure 9a) shows most checkins are located on intersections in the urban road network. These participatory sensing information is nontrivial for marking popular route choices and for identifying congested intersections and road segments where traffic conditions should be improved significantly.

Figure 10 shows the temporal patterns of the checkins. We can clearly see that checkins mainly happened during weekdays, especially during the morning (around 8AM) and afternoon (around 5PM), reflecting the heavy traffic time during the commuting periods.

5.3 A Tale of Two Zones

Our controlled region spans some adjacent livehoods (Cranshaw et al., 2012) with different life activity patterns, including East Liberty and Bakery Square. Figure 11 gives two word clouds generated using the checkin comments in East Liberty and Bakery Square. Users
Figure 10: Temporal Checkin Patterns for the “Traffic” Topic.

Figure 11: Word Clouds based on the Checkin Comments in East Liberty and Bakery Square.

talked about some major stores (e.g., Target and Whole Foods), and restaurants/bars (e.g., BRGR and Kelly) in East Liberty, while they mentioned some professional places (e.g., TechShop and Google) and LA fitness (LAF) in Bakery Square.

We choose two zones, $Z_1$ in East Liberty and $Z_2$ in Bakery Square as shown in Figure 1b, for a closer observation. For the two zones, accurate departure vehicle flows can be detected at the side-street exits of the intersections $F$ and $H$, and of the intersections $O$ and $M$, respectively.

Figure 12 shows the seasonal average vehicle flow patterns for zones $Z_1$ and $Z_2$. In the zone $Z_1$, the traffic flow in weekends is significantly higher than that in weekdays. While in the zone $Z_2$, the traffic flow in weekends is significantly lower than that in weekdays. As we know, the zones $Z_1$ and $Z_2$ respectively contain a major store (Target) and a company (Google), the two figures therefore are consistent with the the common human behaviors of shopping in weekends but working in weekdays.

In the participatory sensing data, there are 2390 checkins of 1054 users in Zone $Z_1$, and 5349 checkins 1260 users in Zone $Z_2$. There are only 253 users appeared in both zones.
Figure 12: Seasonal Average Vehicle Flow Patterns for Two Zones Z1 and Z2.

Figure 13: Checkin and Transition Patterns for Two Zones Z1 and Z2.
We further find the transitions of users for the zones. For each zone, each user transition is defined as two consecutive checkins, one is within the zone and another is outside of the zone, made by a user in a time threshold (four hours in this paper). If the later checkin is in the zone, the transition is an in-zone transition, otherwise it is an out-zone transition. For each zone, the in-zone and out-zone transitions are more related to arrival and departure traffic. The transitions also provide origin and destination (O-D) information for each zone.

Figure 13 gives the checkin and transition patterns for the two zones. In Z1, most checkins have turned to be transitions, which might due to that users do not stay too long after the shopping. In Z1, the number of transitions is significantly lower than that of checkins, as users might stay long and send checkins in their work places. For weekdays and weekends, the
temporal checkin patterns are roughly similar to that of traffic flow patterns. The numbers of both checkins and transitions are still not large enough to support a correlation analysis in statistics with the traffic flow data shown in Figure 12. But the situation should be improved in the near future based on the rapid increasing trend as shown in Figure 3a.

Figure 14 gives the user transition information for zones Z1 and Z2. The origin/destination (O-D) checkins of transitions are distributed broadly in the participatory sensing region. As shown in 14c and 14d, the O-D locations are highly clustered, and the majority of them are covered by a few clusters of sources. For the two zones, they share most of the O-D sources (clusters in blue), although they have an apparent difference in both the checkin and flow patterns. It implies that urban traffic might be mostly generated among a few clustered sources.

The O-D information might be used to further understand the traffic demands, based on some modeling frameworks, e.g., the gravity model (Yang, Jin, Wan, Li, & Ran, 2014) or the radiant model (Simini et al., 2012). Together with popular route information (Wei, Zheng, & Peng, 2012), the information beyond the traffic control region might be useful for recommending time-sensitive alternative routes (Hsieh, Li, & Lin, 2012) to help reducing traffic congestion within urban traffic control systems (Xie, Feng, Smith, & Head, 2014), especially if traffic anomaly (Pan, Demiryurek, Shahabi, & Gupta, 2013) has been quickly identified or predicted. It is also possible to provide carpooling recommendation for users based on the similarity in their O-D transition patterns.

6 Conclusions

In this study, we worked on exploring the potentials of combining physical and participatory sensing data in urban mobility networks. We first presented the basic spatial and temporal characteristics of the sensing data. Basic human mobility patterns were then extracted directly based on user checkins. For users, the entropy and time-dependent regularity were further disclosed by clustering their visitation based on geo-locations to understand the degree of predictability. Some initial results were presented for illustrating the usages of the sensing data in urban mobility applications, e.g., accurate timing and reasoning an anomaly detection, unveiling nontrivial traffic-related information in topic-specific checkins, and revealing origin and destination patterns based on transitions between user checkins. Our study might shed some lights on further studies for improving urban mobility.

Acknowledgements. This research was supported in part by the T-SET University Transportation Center at Carnegie Mellon University, and the CMU Robotics Institute.
References


Hsieh, H.-P., Li, C.-T., & Lin, S.-D. (2012). Exploiting large-scale check-in data to rec-


demand estimation using location based social networking data. In *Transportation Research Board Annual Meeting*. Washington, DC.
Mapping Urban Soundscapes

Tae Hong Park

keywords: sound-mapping, soundscape, noise pollution

Department of Music & Performing Arts Professions
Steinhardt School
New York University 35 West 4th Street
Suite 1077
New York, NY 10012

thp1@nyu.edu
504-909-2267
Abstract
The study and practice of environmental cartography has a long and rich human history with early examples found in ancient 5th century Babylonian and ancient Chinese cultures (Horowitz, 1988; Miyajima, 1998). More recently, with the advent of precise measuring techniques enabled via global positioning systems, environmental sensors, cloud computing, and various components of geographic information systems, digital cartography has become ubiquitous and perhaps even a necessity for the modern global citizen. As has been the case throughout history, mapping systems generally still follows the traditional mapping model described via fixed landmarks such as coastlines, buildings, parks, and roads. However, human interaction with the environment is immersive, demanding a poly-sensory engagement where all but one of our senses is visually oriented. Even in contemporary mapping practices, invisible environmental dimensions are severely underrepresented partly due to: (1) importance of the ocular reflected in the so-called eye culture (Berendt, 1988) and (2) technical difficulties in addressing spatio-temporality needed to capture non-ocular environmental dimensions. In 2011, the Citygram Project was launched to explore non-ocular spatio-temporal energies through strategies that address the collection, mapping, analysis, and archival of invisible spatial energies. The project’s multi-iteration structure is currently in its first iteration and focuses on acoustic energy. The aim is to explore, research, and engineer infrastructural frameworks to contribute to existing mapping paradigms by addressing critical components necessary for capturing soundscapes. As such, spatio-acoustic energy is captured via a flexible sensor network, which is then analyzed, visualized, and mapped. The project’s goal also includes advancing multimodal geospatial research by embracing the idea of time-variant, poly-sensory cartography to better understand ecological questions. In this paper we summarize efforts in developing concepts, technologies, and analysis techniques that render data-driven multi-format maps with an overarching aim to better understand of our environment and the often-ignored byproduct of urban environments – noise pollution.

1. Introduction: Soundscapes and Noise
Humans have shown a remarkable ability to adjust to changing environments and studies suggest that we “have undergone rampant adaptation” in the last 200,000 years of history (JJ Cai, Macpherson, Sella, & Petrov, 2009). In the last 200 years, however, the size of cities, population growth, and accompanying urban infrastructural complexities, along with its multimodal byproducts, have reached astonishing numbers. The industrial revolution, in particular, has been a cataclysmic contributor to rapid worldwide population growth and change in our natural environment (Lutz, Sanderson, & Scherbov, 2008), and this includes soundscapes (Schafer, 1993; Wrightson, 2000). Modern city-dwellers are all too familiar with the constant cacophony of urban machinery and the ubiquity of myriads of noise pollutants, regardless of time and space. For New Yorkers, the city’s noisy soundscapes have become second nature. Adapting to noise pollution, however, comes with serious associated health risks and according to Bronzaft, one of the leading experts in environmental psychology, “It means you’ve adapted to the noise … you’re using energy to cope with the situation. That’s wear and tear on your body” (“Arline Bronzaft Seeks a Less Noisy New York - NYTimes.com,” n.d.). Studies show that such “wear and tear” does not just contribute to hearing impairment, but also non-auditory health risks, including adverse effects on children’s learning skills, hypertension, and sleep deprivation, as well as gastrointestinal, cardiovascular, and other physiological disorders (A. L. Bronzaft, 2002; Gary W Evans, Lercher, Meis, Ising, & Kofler, 2001; Kryter & others, 1970; Lang, Fouriaud, & Jacquinet-Salord, 1992; Passchier-Vermeer & Passchier, 2000; Van Dijk, Souman, & De Vries, 1987; Ward, 1987; Woolner & Hall, 2010; Zhao, Zhang, Selvin, & Spear, 1991), (Barregard, Bonde, & Öhrström, 2009; Jarup & Babisch, 2008; Knipschild, 1977). This notion of human “adaptation” is especially concerning if we consider how little adaptation time we have had since the expansion of
manmade urban environments. There is a reasonable case to be made that the current noise pollution situation will significantly worsen with rapid population growth, which in turn will likely contribute to the expansion of ever denser and larger megacities worldwide: by 2050, it is projected that 3/5 of the global population will live in one of these megacities. Even as recently as the 1970s, noise was considered a mere nuisance. During this period, however, environmental researchers also began to uncover its negative health impact on city dwellers (Arline Bronzaft, n.d.; A. L. Bronzaft, 1981; A. Bronzaft & Van Ryzin, 2007; G. W. Evans, Hygge, & Bullinger, 1995; Gary W Evans & Lepore, 1993). In 1972, the US Congress passed the Noise Pollution and Abatement Act, which led to the developed of noise emission standards for motor vehicles, aircraft, and industrial equipment; and in 1975, one of the world’s first comprehensive noise code was developed in Portland, Oregon. Like many noise codes, the Portland code defines noise using a spatio-temporal schema and distinguishes sound from noise via sound level, location, and time. This reframed noise to be thought of as sound that exceeded a certain sound pressure level (SPL) threshold further articulated in terms of space and time. New York City (NYC) has been particularly sensitive to its noisy soundscape, and for good reason: since 2003 more than 3.1 million noise complaints have been logged by NYC’s 311 city service hotline1 representing the top category of complaints as quantified by the 311 reporting mechanism. Other cities nationwide that have implemented 311-style citizen hotlines have figures comparable to NYC: recent consumer ranking of the noisiest cities in the United States include Chicago, Atlanta, Philadelphia, San Francisco, and Houston (Washington, n.d.). Noisy urban environments – something that acoustic ecologist Schaefer refers to as lo-fi soundscapes (Schafer, 1977) – is unsurprisingly an international phenomenon and continues to be one of the main environmental problems facing Europe today. For example, studies in the United Kingdom have shown that the general population lives above WHO noise level recommendations (WHO Guidelines for Community Noise) where an increase of noise has been recorded between 1990 – 2000 (Skinner & Grimwood, 2005). In another study, it has been shown European Union households willingness to annually pay up to 34 Euros per decibel of noise reduction (Barreiro, Sánchez, & Viladrich-Grau *, 2005; Navrud, 2000).

Although we have come a long way since recognizing that noise is not just a mere nuisance or irritation (Skinner & Grimwood, 2005) for humans, noise codes as written and enforced today are problematic in several respects:

1. The metrics by which noise is defined are based on definitions of excessive “volume” that are either severely subjective or, when standard SPL measurements are used, fail to reflect how sound is perceived. For example, soothing ocean waves at 80 dB and the sound of blackboard fingernail scratching at the same level are not perceived in the same way.

2. Cities’ capacity to effectively monitor, quantify, and evaluate urban noise is very limited.

3. The mechanism for noise enforcement is impractical as noise is fleeting in nature: even when law enforcement officers do make it to a reported “noise scene,” chances are that any noise pollution traces will have disappeared completely by the time they arrive.

4. Noise complaints are typically reported via 311 hotlines or directly reported to the police. However, these tools are inadequate for reporting or combating noise. For example, studies show that only 10% of surveyed residents who were experiencing noise issues bothered to contact authorities: most directly confronted the person responsible (Skinner & Grimwood, 2005).

2005) which may partly explain the 4.5 annual noise-related killings affiliated with neighbor disputes (Slapper, 1996).

With the recent maturing of cost-effective technologies including wireless communication networks, cloud computing, crowd-sourcing/citizen-science practices, and the explosion of Big Data science as a growing research field, the past few years has provided an opportunity to re-examine many of the issues pertinent to capturing soundscapes – in particular noise pollution – by creating a comprehensive real-time and interactive cyber-physical system (CPS) for collecting, analyzing, mapping, and archiving soundscapes. Additionally, considering the increasing willingness of cities to provide public access to spatial data2 and integrate data science techniques and civic participation towards public policy-making decisions (Dickinson et al., 2012), an even more compelling case for developing an adaptive, scalable, and comprehensive CPS system for mapping our hyperdimensional environment can be made. In the following sections we will provide an overview, which aims to begin to address many of the issues that exist in creating soundmaps today.

2. The Citygram Project: History and Overview
The Citygram Project (T. H. Park et al., 2013; T. H. Park et al. 2014; T. H. Park et al., 2012; T. H. Park et al. 2014; T. H. Park et al., 2014) began to take shape in 2011 with the observation that past and present digital cartographic paradigms are typically based on static landscapes characterized by slowly changing landmarks such as buildings, avenues, train tracks, lakes, forests, and other visible objects. Ocular, physical objects, although critical in any mapping model, are not the only elements that define environments; various energy types including non-ocular, acoustic energies also define and characterize the environment and our cities. Noticing the underrepresentation of sound in modern interactive mapping practices, we began to explore and develop concepts and ideas to enable spatio-temporal mapping via real-time capture, streaming, analysis, and human-computer interaction technologies. In essence, initial research focused on examining the feasibility of creating “soundmaps.” Although Citygram began as “sound-mapping” project in the early stages, the idea of an iterative interactive “non-ocular” mapping concept based on data-driven paradigms quickly followed suit as a number of fundamental shortcomings in modern digital mapping systems were observed: (1) underrepresentation of “non-ocular” energies, (2) spatio-temporality, and (3) spatio-temporal granularity. Google Maps, for example, updates spatial images every one to three years3 reflecting the low sampling rate needed to capture the nature of slowly changing landscapes, which is clearly inadequate for soundmaps. Soundmaps require in the order of tens of thousand samples per second for any meaningful representation. Citygram thus began as an effort to augment and contribute to next-generation geospatial research by embracing the idea of real-time, non-ocular, and poly-sensory cartography, ultimately enabling an immersive and richer representation of the dynamicity of physical environments. Our first iteration, Citygram One (T. H. Park et al., 2012), we focus on exploring spatio-temporal acoustic energy via soundmaps to reveal meaningful information including spatial loudness, noise pollution, traffic patterns, and spatial emotion/mood enabled by a comprehensive cyber-physical system that includes a robust sensor network, server technology, visualizations, interaction technologies, and machine learning techniques.

2 https://nycopendata.socrata.com/
3 https://sites.google.com/site/earthhowdoi/Home/ageandclarityofimagery
Citygram’s data-driven maps are based on spatial quantitative low-level feature vector data streams processed by remote sensing devices (RSD) and stored on our server. The project has grown since its inception in 2011 with early collaborators from California Institute of the Arts (CalArts) and more recently adding researchers from NYU Steinhardt School, NYU’s Center for Urban Science and Progress (CUSP), and NYU’s Interactive Telecommunication Program (ITP). Through the NYU CUSP collaboration and its Sound Project in 2013, the focus has further narrowed in scope leading to concentration in urban noise exploration. An early proof-of-concept heatmap visualization of spatio-temporal acoustic energy is shown in Figure 1. The dynamic heatmaps is overlaid on a standard Google Maps API.

![Figure 1. Citygram dB\textsubscript{RMS} visualization](image.png)

2.1 Related Work
One of the earliest works in capturing and analyzing environmental sound is attributed to research conducted by R. Murray Schafer. Schafer is credited with coining the term soundscape, and is one of the founders of the World Soundscape Project\(^4\) (WSP). WSP is concerned with education and research in acoustic ecology which include concepts such as lo-fi and hi-fi (Schafer, 1977) sounds as it relates to low and high signal-to-noise ratio (SNR) sound environments. Work in soundscape has been traditionally in the realm of capturing environmental sounds while simultaneously trying to bring attention to disappearing soundscapes largely due to human environmental intervention accelerated by the industrial revolution (Wrightson, 2000). The group engaged in numerous projects that primarily entailed capturing sound environments by recording specific soundscapes in a variety of locations using portable recording devices. Their efforts lead to published recordings such as *The Vancouver Soundscapes* (1973), *Five Village Soundscapes* (1975), and *Soundscape Documentation DVD-ROM* (2009).

In recent years, a variety of soundscape-like projects have begun to emerge, many touching upon principles pioneered by Schaefer. In particular, with the maturation of affordable hardware and software technologies, smaller portable devices often in the form of smartphones or other handheld devices, and the explosion of personal wireless telecommunication as vital accessory of the *homo urbanus*, numerous application workflows have been explored to capture and measure soundscapes and environmental noise. Many of these applications tap into a somewhat new phenomenon of the “citizen-scientist” to help contribute in solving problems that are made difficult by addressing issues of spatial data granularity. In 2008, an application called *NoiseTube* was developed at Sony Computer Science Laboratory to measure noise via smartphones, enabling the development of *crowd-sourced* noise maps of urban areas (Maisonneuve, Stevens, & Steels, 2009). *NoiseTube* attempted to address the “lack of public involvement in the management of the commons” by empowering the public with smartphones to measure “personal” noise pollution

\(^4\) http://www.sfu.ca/~truax/wsp.html
exposure via mean dB SPL levels. *WideNoise*, a 2009 Android/iOS application is also a citizen-science noise metering example that includes an active world map of current noise levels (“widenoise @ cs.everyaware.eu,” n.d.). *WideNoise* includes a social user experience component, which encourages active participation and also includes a sound sample-tagging feature that allows users to annotate sounds by associating it with its source ID and mood labels. Another related project is *Motivity* (2010), which employs a small number of stationary decibel meters at key intersections in the Tenderloin neighborhood of San Francisco. Developed as an acoustic ecology project to demonstrate the efficacy of noise metering in a high-traffic area, the project uses an instrumentation system consisting of fixed microphones with embedded computing systems placed at intersections within a 25-block area. As with the other projects, *TenderNoise* and *WideNoise* both use the one-size-fits-all SPL metric to evaluate noise. In the area of preservation and capturing dying soundscapes such as rainforests, *Global Soundscape* is a project from Purdue University that offers simple tagging options and additional verbose descriptors inputted via a custom smartphone application. One of the goals of this project was to collect over one million natural soundscapes as part of a special *Earth Day* experience on April 22, 2014. Like many of the other software solutions, *Global Soundscape* also provides a mapping interface with nodes that represent locative audio snapshots contributed by citizen-scientists. The *Locustream SoundMap* is another soundscape-based project and is based on a so-called “networked open mic” streaming concept. In essence, Locustream aims to broadcast site-specific, unmodified audio through an Internet mapping interface by participants referred to as “streamers.” Streamers are persons who deploy custom-made *Locustream* devices, which are provided from the developers to potential users in order to share “non-spectacular or non-event based quality of the streams.” This project was one of the many sources of inspiration for our project and we have taken this concept a step further by providing means for anyone with a computer and microphone to participate in serving as a “streamer” as further discussed in Section 2.2. Other examples include remote environmental sensing such as *NoiseMap* (Schweizer, Bärtl, Schulz, Probst, & Mühlhäuser, 2011) and *Tmote Invent* (So-In et al., 2012) which utilize a periodic record-and-upload SPL level strategy for noise monitoring. A final example is a project from the Netherlands called *Sensor City* (Steele, Krijnders, & Guatavino, 2013). The project aims to deploy “hundreds” of fixed sensors equipped with high-end, calibrated acoustic monitoring hardware and its own dedicated fiber-optic network around a small city. *Sensor City* aims to explore human perception and evaluation of acoustic environments within urban settings to qualify soundscapes via machine learning techniques.

In the following sections we provide an overview and summary of the various components of the Citygram Project that form a cyber-physical system for capturing soundscapes.

### 2.2 Sensor Network and Remote Sensing Devices: Fixed and Crowd-Sourced

Creating an immersive soundmap begins with capturing spatio-temporal sound. That is, capturing sound in real-time via an acoustic sensor network. Our sensor network design philosophy is based on adopting robust, cost-effective, and flexible remote sensing devices (RSD) that communicate through cloud-computing technologies to create a dense sensor network infrastructure. This includes addressing issues associated with traditional spatially sparse monitoring practices that cover large areas with a small number of bulky and often costly sensors. These designs have the advantage of very high sound quality, but at the same time suffer in the area of scalability. Our strategy also aims to address concerns related to an overreliance on consumer handheld devices (e.g. smartphones and tablets) for sensor network creation. These designs may have adverse

---

5 http://tendernoise.movity.com
6 https://www.globalsoundscape.org/
effects on data quality due to inadequate onboard hardware components as well as issues related to calibration, control, and recording variability in personal crowd-sourced sound capture practices. Our sensor network strategy aims to create a dense sensor networks through rethinking of the functionality, utility, ubiquity, and adaptation of computing platforms in order to render seamless server-RSD interoperability via fixed and crowd-sourced environmental sensing paradigms. Figure 2 shows the Citygram sensor network infrastructure with a server and various forms of RSDs including desktop and laptop computers; handheld devices such as smartphones and tablets, and fixed, calibrated RSDs as further described in the following sections.

Figure 2. Sensor network infrastructure

Figure 3. (a) Android fixed RSD proof-of-concept showing processor and ADC (left-hand) and (b) single MEMS (right-hand)

Fixed RSDs Fixed RSDs are permanently installed in “fixed” locations to provide consistent, reliable, secure, and calibrated audio data to our server. These RSDs, which constitute a distributed computing network, use identical hardware and software components to ensure data consistency. To date, a number of initial systematic tests have been conducted to select suitable components for RSD development. Tests have included consideration of audio capture capability, processing power, RAM, onboard storage, OS flexibility, wireless connectivity, power consumption, I/O expandability, robustness/sturdiness, cost-effectiveness, and technology transferability. A number of RSD candidates were considered including Alix boards, Raspberry Pi hardware, Arduino microcontrollers, and a variety of consumer handheld devices. Our analyses efforts have led to adopting the Android mini-PC platform for our system by considering the above factors as well hardware footprint: the mini-PC is approximately the size of a jump-drive as
shown in Figure 3(a). Additional measurements and analyses workflows have led to identifying potential microphones by considering frequency response, pickup patterns, dynamic range, size, power consumption, and durability. Our current proof-of-concept RSD currently includes a single custom Micro-Electro-Mechanical Systems (MEMS) microphone board as shown in Figure 3(b). Since April 2014, we have been conducting field tests via deploying a number of RSD nodes in normal outdoor weather conditions in the Brooklyn area. These low-cost, fixed RSDs capture, analyze, and transmit consistent soundscape reporting via distributed and cloud computing client-server architectures.

**RSD Software Update** Secure remote software updates for our RSDs are essential in allowing for efficient management and development of our sensor network. As further discussed below, for our mobile/crowd-sourced RSDs, updates can be accomplished manually by users via download links or built-in auto-update mechanisms in the case of registered mobile applications. Our fixed RSDs, however, use a custom software update module to enable remote sensor network management, as manually updating large number of deployed RSD is not only impractical and costly but also detrimental for system scalability.

**Crowd-Sourced RSD** Our crowd-sourced RSDs are based on a design philosophy of plug-and-sense whereby any computing device with a microphone and Internet connection can be rendered into an RSD and allow users to become streamers. This includes smartphones, tablets, “phablets” (phone-tablet hybrids), laptops, and desktop computers running “apps” and other software add-ons that run on popular commercial software. We believe that our hybrid system exploits the benefits of what both fixed and crowd-sourced RSDs have to offer in facilitating the creation of a dense sensor network to produce high level of spatial granularity. Spatial granularity is critical in addressing the unpredictability of environments such as urban spaces. Unpredictability, by its very definition, is problematic when aiming to address spatial sensor scarcity distributions\(^7\) through modeling strategies in order to fill the spatial “data gaps” (Aberer et al., 2010; Saukh, Hasenfratz, Noori, Ulrich, & Thiele, n.d.) (e.g. a commonly used method is the generalized additive model (GAM) (Ramsay, Burnett, & Krewski, 2003) based on statistics). A number of prototype software have been developed for Android and desktop platforms (T. H. Park et al., 2013; T. H. Park et al., 2014). These crowd-sourced RSDs are designed to capture, analyze, and stream audio data including feature vectors in addition to our fixed RSDs to facilitate the creation of a dense network while inviting meaningful community and citizen-science participation.

**2.2.1 Sensor Deployment**
Our sensor deployment strategy follows a multi-stage procedure based on incremental deployment steps. A small-scale sensor network step is currently being tested for seamless end-to-end functionality of physical and virtual components. This includes performance validation of fixed RSD nodes, mobile/crowd-sourced RSDs, server-RSD data communication reliability, database query/update efficiency, and data visualization. As such, we are currently testing a small number of RSD deployed outside the “lab” environment. Our long-term and large-scale deployment plans include activating citizen-scientists and adapting our CPS to existing urban infrastructures. This includes the application of Citygram for artistic purposes including real-time music performance and composition, real-time data-driven visualization, and the development of interactive tools. One key future deployment strategy includes partnering with non-commercial (e.g. NYC currently has 59 unlimited free hotspots) and private sector

\(^7\) e.g. NABEL and OstLuft have a combined sensor node count of five for the entire city of Zurich
organizations, which have taken initiatives to provide free and open Wi-Fi to urban city dwellers. For example, in 2013 Google sponsored the creation of free Wi-Fi to 2000+ residents, 5000+ student populations, and hundreds of workers in Manhattan’s Chelsea area. Another example is NYC’s initiatives to “reinvent” 11,412 public payphones⁸. These payphones produce approximately $17.5 million annual revenue primarily from advertising but its function as a public telecommunication station has practically been rendered obsolete – it has been reported that roughly 1/3 of the payphones in Manhattan are inoperable as communication apparatuses⁹. NYC’s recent call-for-proposals to “reinvent payphones” aims to install, operate, and maintain up to 10,000 public payphone nodes with free Wi-Fi and other technologies. The urban payphone infrastructure could serve as an ideal large-scale deployment mechanism for our CPS sensor network as each station will provide uninterrupted power supply, data communication capability, (additional) weather protection, and as whole, provide an opportunity to repurpose existing (but practically obsolete) urban infrastructures for large-scale fixed RSD deployment. Such a model would be straightforwardly transferable to other cities around the world.

2.3 Machine Learning and Analytics: Soundscape Information Retrieval (SIR) The Big Data analysis portion of our research is a critical research component and involves the fundamental understanding of soundscapes through descriptors including semantic, emotive, and acoustic descriptors. In the case of semantic soundscape descriptors, much of the work can be found in the realm of acoustic ecology (Schafer, 1977) where the identity of the sound source, the notion of signal (foreground), keynote (background), soundmarks (symbolically important), geophony (natural), biophony (biological), and anthrophony (human-generated) play important roles. However, as our immediate focus lies in automatic classification of urban noise polluting agents, a first step towards automatic classification is the development of agreed-upon urban soundscape taxonomy. That is, (1) determining what sound classes occupy urban soundscapes, (2) developing an agreed-upon soundscape namespace, and (3) establishing organizational and relational insights of its classes. This research, however, is still underexplored and a standardized taxonomy is yet to be established (Brown, Kang, & Gjestland, 2011; Guastavino, 2007; Marcell, Borella, Greene, Kerr, & Rogers, 2000). In an effort to develop urban environmental noise taxonomy, we are currently developing software tools to conduct surveys using custom interactive online annotation software. Our methodology follows an open-ended labeling paradigm similar to (Marcell et al., 2000). We also are also developing a urban soundscape taxonomy that more accurately reflects the notion of “collective listening” rather than relying on the opinions of a few (K. Foale, n.d.). We believe that our methodology of inviting both researchers and the public to define and refine the pool of semantic sonic concepts has the potential to contribute to pluralistic soundscape taxonomy. Furthermore, using crowd-sourcing techniques for taxonomical development may yield more than an expanded tag-pool for labeling audio events as it has the potential of revealing connectivity between sounds and everyday concepts as defined by a global community. As such, we are undertaking steps to explore semantic mining techniques to reveal fundamental insights coded in subjective, qualitative associations between sounds and concepts that can be quantified and rendered into a subjectively communicable format. Of critical importance in this approach is to ensure a sufficiently large collection of data. We are thus developing software to scrape data from the Internet using keyword search strategies in combination with natural language processing (NLP) techniques to reveal semantic relevancy and patterns. As an initial proof-of-concept we examined urban soundscape mining through data

obtained Freesound\textsuperscript{10}. Although Freesound is a rich resource for annotated audio files, initial semantic analysis proved to be tricky due to the amount of noise (unrelated words) present in the data. However, after employing a number of de-noising filters including tag normalization, spelling correction, lemmatization (Jurafsky & Martin, 2009), and histogram pruning, we were able to substantially clean the 2,203 raw tags to 230 tags obtained from 1,188 annotated audio files. An analysis of the filtered tags suggested that with additional filtering techniques we could gain insights into hierarchical and taxonomical information in addition to our current conditional probability techniques. For example, many tags could be categorized as geophones, biophones, or vehicular sounds suggesting the potential for applying NLP-based semantic distances available in open-source software packages such as \textsc{WordNet} (Fellbaum, 1998).

In addition to semantic descriptor analysis, we are currently investigating research efficacy in spatio-acoustic \textit{affective} computing (Barrett, 1998). We believe that these supplemental emotive parameters will further provide insights into perceptual qualities of noise pollution, its semantic associations, and soundscape perception in general while also help in determining human affective states with respect to urban sounds.

The final area of analysis research is in low-level acoustic descriptors. These \textit{quantifiable} audio descriptors (Keim, Panse, & Sips, 2004; Lerch, 2012; Müller, 2007; Peeters, 2004) are extracted every few hundred milliseconds, commonly subjected to statistical summaries (Aucouturier, 2006; Cowling & Sitte, 2003; Meng, Ahrendt, Larsen, & Hansen, 2007), and then fed to a classifier as further discussed in the next section. Standard features for representing audio include centroid (SF), spread (SS), flatness measure (SFM), spread, flux (SF), Mel-Frequency Cepstral Coefficients (MFCC) in the frequency-domain; attack time, amplitude envelope, Linear Predictive Coding (LPC) coefficients, and zero-crossing rate, in the time-domain. Before classification, it is also helpful to conduct feature dimensionality reduction for efficient and effective feature representation. In adding quantitative soundscape analysis techniques, our aim is to develop a comprehensive analysis framework whereby semantic, emotive, and quantitative soundscape dimensions are included. This will ultimately inform us in developing our automatic urban soundscape classifiers. It should, however, be noted that noise is not entirely subjective and that feature vectors are not invariant. In some cases, depending on environmental conditions, a sound source may not always be perceived the same way, and a sound’s feature vectors may also change for the same sound class. Sound class variance may also be affected by the notion of acoustic event \textit{presence} in the form of foreground, middle-ground, and background. In a sense, it could be argued that traditional noise measurement practices primarily focus on the foreground characteristic of noise quantified via dB levels – sounds that were produced in the background would yield low noise rankings and thus unnoticeable although it may in reality cause irritation to listeners nearby. To systematically determine the effect of \textit{presence}, we are including a foreground – background field in our crowd-sourced annotation workflow that describe each acoustic event tagged by the annotator.

\subsection*{2.3.1 Classification}

The field of automatic soundscape classification is still in its nascent stages partly due to a number of factors including: (1) the lack of ground truth datasets (Giannoulis et al., 2013), (2) the underexplored state of soundscape namespace, (3) the overwhelming emphasis on speech recognition (Gygi, 2001; Tur & Stolcke, 2007; Valero Gonzalez & Alias Pujol, 2013), and (4) the sonic complexity/diversity of soundscape classes. A soundscape can literally contain any sound, making the sound classification task fundamentally difficult (Duan, Zhang, Roe, & Towsey, 2012). That is not to say that research in this field – something we refer to as Soundscape

\footnote{\textsuperscript{10} Crowd-sourced online sound repository contains user-specified metadata, including tags and labels.}
Information Retrieval (SIR) – is inactive: research publications related to music, speech, and environmental sound as a whole has increased more than four-fold between 2003 and 2010 (Valero Gonzalez & Aliás Pujol, 2013); and numerous research subfields exist today, including projects related to monitoring bird species, traffic, and gunshot detection (Jinhai Cai, Ee, Pham, Roe, & Zhang, 2007; Clavel, Ehrette, & Richard, 2005; Mogi & Kasai, 2012; van der Merwe & Jordaan, 2013).

One of the notable initiatives in SIR research began recently in 2013 with the creation of the IEEE D-CASE Challenge for Computational Audio Scene Analysis (CASA) (Giannoulis et al., 2013). Although training and evaluation of SIR systems were primarily focused on indoor office sounds\textsuperscript{11}, it is still worthwhile to note some of the SIR techniques presented at the Challenge. In the area of feature extraction, MFCCs were widely used, although in some studies, a case was made for time-domain and computer vision approaches realized via matching pursuit (MP) and a k-NN-based spectrogram image feature (Dennis, 2011). The former used a dictionary of atoms for feature presentation (Chu, Narayanan, & Kuo, 2008, 2009) and the latter exploited spectrogram images for acoustic event detection (AED) and acoustic event classification (AEC). Both methods were demonstrated as alternative methods to MFCCs and showed robust performance in the presence of background noise. Some of the classifiers that were omnipresent included k-NNs, GMMs, SVMs, HMMs, and SOFMs based on expert-engineered feature vectors also reported in (Duan et al., 2012). The road ahead, however, is still largely uncharted and fundamental topics concerning the taxonomy and soundscape semantics, soundscape dataset availability, and the development of comprehensive and robust classification models that are adaptable to the diversity, dynamicity, and sonic uncertainty of outdoor soundscapes still remains challenging (Peltonen, Tuomi, Klapuri, Huopaniemi, & Sorsa, 2002).

To address taxonomical and semantic side of soundscape research, we are currently developing crowd-sourced annotation tools to collect tags, labels, and soundscape descriptions through semantic data mining techniques. This has dual functionality of gaining insights into the soundscape namespace and also collecting ground truth data for machine learning. The latter research component entails crowd-sourcing multi-person tagged acoustic events in collaboration with various international universities. The database is projected to contain multiple annotations per sound class. The exact number of sound classes will be determined after careful analysis of tags/labels and community-provided noise complaint reports as further discussed below. The prototype annotation software consists of standard playback controls, zoom functionalty, and tagging interfaces, including text entry for open-ended tagging, sound ID, start and end markers for each acoustic event, input for valence/arousal amount, and event fore/middle/background. Once our first phase tagging efforts are complete, we expect the creation of a rich dataset of ground truth data to enable our urban noise classification research.

One of the key issues in SIR is its sonic, spatial, and temporal diversity. These factors make SIR-based machine learning fundamentally difficult. The aim at this stage, however, is not to solve the urban SIR problem per se. Rather, the goal is to develop automatic urban sound classification algorithms that can detect and classify the most “popular” (unpopular?) noise pollutants in cities like NYC; benchmark classification performance, and progressively improve and expand soundscape class identification. And although developing a comprehensive soundscape classifier with large number classes is the ultimate goal, if we can identify a smaller but significantly impacting subset of noise pollutants as determined by city-dwellers, the problem then becomes more manageable and a iterative procedure can be applied. Thus, for the classification portion we are focusing on (1) classifying some of most “popular” noise polluting agents and (2) strategically increasing the collection of noise classes guided by a sound class

\textsuperscript{11} dissimilar to music and speech sounds although arguably a type of “environmental sound”
priority scheme, based on crowd-sourced *noise agent rankings*. To get preliminary assessment of this notion of *noise agent ranking*, we have analyzed the NYC 311 noise complaint dataset which shows the following class distributions representing four years of data: 54% complaints included words *car* or *truck*, 49% *music*, 20% *people* or *talking*, 14% *construction*, and 10% the word *dog*. In other words, if we focus our attention to a smaller subset of soundscape classes (those that are most “popular”) and expand our algorithms to automatically recognize classes in an incremental methodology (include less “popular” ones), then the classification task can be divided into a number of iterations that are more manageable. In addition to the 311 dataset we aim to analyze other similar datasets (“AMERICA’S NOISIEST CITIES - Tags: CITIES & towns -- Ratings & rankings NOISE pollution,” n.d.) from cities like Chicago, Atlanta, Philadelphia, San Francisco, and Houston to validate the efficacy of such a noise ranking system – in the European Union, for example, road traffic noise accounts for 32% of noise events that are above 55 dB(A) (Barreiro et al., 2005). With the amalgamation of fundamental knowledge gained in the analysis research part, we aim to transition from a position of questioning, – “what is noise?” – to a position of enunciation – “this is noise.”

As a preliminary step in soundscape classification we have conducted a number of experiments including acoustic event detection (AED) (T. H. Park, et al., 2014). The overall goal for this initial research was to explore the SIR feature space to inform subsequent classification research strategies. The majority of audio classification systems combine AED and AEC. AEC (here, background is simply considered to be another class) has been proven especially effective when an audio scene is highly specific, its event classes well defined, and small in number. In our research, we are exploring an approach where AED is conducted separately from AEC. That is, we first do AED and only when an acoustic event is detected, do we classify the sound. This is due to a number of strategic reasons including system efficiency and sensor network transmission bandwidth. Ideally one can easily imagine transmission of sound IDs in the form of integer numbers: for 100 sound types we would only need to 8 bits to be transmitted for identifying the source. To address system efficiency (computation load, energy efficiency, and bandwidth usage), data from an RSD to a server is only transmitted when an acoustic event is detected. The latter point considers the fact that soundscapes vary in activity depending on location and time, and continually running AEC 24/7 is therefore inefficient as acoustic scenes will often be inactive and uneventful.

Our current AED algorithm uses 19 feature vectors including RMS, ZCR, SF, SFM, SF, SS, and 13 MFCC coefficients. We designed and tested two AED systems where in one we employ an adaptive noise removal algorithm (T. H. Park, Lee, et al., 2014). This helps in adaptively and dynamically attenuating a soundscape’s background noise, which is often in flux throughout the day. This in turn improves feature computation due to background noise removal. Another AED algorithm we employed linear discriminant analysis (LDA) to classify the input into two classes: event or non-event. Both resulted in approximately 73% correct classification of 176 acoustic events totaling 76 minutes in duration. The most salient features for our preliminary AED tests were ranked as follows: SS, SFM, and MFCC. Furthermore, we were able to improve performance by approximately 12% by increasing the statistical window size (and “bag-of-fames”) to approximately one second (85% classification). Feature ranking, however, remained similar. Other initial classification research included unsupervised learning techniques, dimensionality reduction (PCA), and k-NN clustering.

A final area of research we are examining is feature learning. This area of machine learning is based on automatic “feature learning” opposed to engineering features “manually” via expert

---

12 www.sf11.org lists barking dog, people talking, and car alarms as top 3 examples for noise complaints.

13 Houston noise code lists vehicles, amplified sound from vehicles, and noise animals are top noise examples.
knowledge. Feature learning requires large quantities of examples. Deep learning and deep networks, as they are also referred to, are based on neural networks, which are particularly well suited for learning complex functions that are difficult to model manually. Preliminary experiments have led to some interesting initial results in (Jacoby, 2014). Although, the feature learning method did not outperform traditional discriminative models (SVM), it showed interesting artifacts. For example, for some urban sound classes (those that exhibited clear resonant characteristics such as pitch), it outperformed discriminative models. One way to look at this result is that, as we are using algorithms, methods, and approaches primarily developed for music and speech, they may work well only for specific soundscape classes.

2.4 Interaction and Exploration
One of the goals of creating a comprehensive cyber-physical system is including mechanisms for interactive exploration. We are thus developing online access and exploration technologies not only for researchers but also for citizen-scientists, students, artists, and the general public. The current proof-of-concept web interface is designed to function as an interactive environmental exploration portal and is built on the Google Maps API. Also, a number of visualization prototypes have been realized (T. H. Park et al., 2013; T. H. Park, Turner, et al., 2014) providing real-time visualizations and accompanying interfaces for standard web browsers. The interface dynamically visualizes RSD-streamed audio data and also provides the ability to animate historical data stored in the server database. The historical data serves as an archival module, which stores low-level spatio-acoustic feature vectors. To enable users to hear the “texture” and characteristics of spaces without compromising private conversations that may be inadvertently captured in public spaces, we employ a custom voice blurring technique based on a granular synthesis (Roads, 1988). To accomplish these conflicting tasks – blurring the audio while retaining the soundscape’s texture – a multi-band signal processing approach has devised as detailed in (T. H. Park, Turner, et al., 2014).

3. Future Work
Citygram is designed to be an iterative, long-term project with a current focuses on soundmapping and acoustic ecology. Completion of iteration one will result in a comprehensive cyber-physical system (CPS) that includes modules for real-time data capture, data streaming, sensor networks, analytics, archiving, and exploration. These same modules will form a comprehensive CPS framework that will facilitate system expansion and inclusion of additional sensors for our RSDs to capture other spatio-temporal data. Our current fixed RSD designs use standalone analog-to-digital audio converters connected to one of the onboard USB I/O ports. Using onboard I/O and wireless Bluetooth communication capabilities will allow for flexibility of adding additional sensors such as accelerometers, air pollution sensors, electro-magnetic, humidity, and other sensors. Iteratively focusing on a particular “non-ocular” spatio-temporal energy type is part of our long-term strategy. Additionally we also plan to fold-in and consider other spatial data modalities including census data, crime statistics, real-estate data, weather, as well as social media network data feeds such as twitter feeds. We anticipate that creating a multi-format, multidimensional, and multimodal CPS will allow for the opportunity to provide immersive, interactive, and real-time maps that will improve our understanding of our cyber-physical society.

4. Summary
We presented a cyber-physical system approach for interactive sound mapping and provided an overview of the Citygram project; summary of various software and hardware prototypes; strategies for capturing spatio-temporal granularity and RSD deployment including discussions related to scalability; soundscape-based machine learning topics including soundscape taxonomy,
namespace, and ground truth data; as well as research and development that is currently underway. We discussed project’s scope including its potential for providing public service by embracing an immersive mapping paradigm. We also discussed its potential contribution in combating urban noise pollution by providing mechanisms to gain insights into spatio-temporal noise pollution patterns as well its potential for evaluating noise agents in the context of the broken windows theory¹⁴ (Harcourt & Ludwig, n.d.) – in our case, in the context of a “broken sound theory.” It has been shown that noise has been symptomatic of criminal activity including confrontations between neighbors; spousal/child abuse and abuse of the elderly; and drug dealers who, in NYC for example, have been pushed from the streets into buildings and often marked by noise such as constant loud music, “customer” visitation noise at all hours of the night, and repeated usage of whistles and other “all-clear” signals (Blair, 1999; A. L. Bronzaft, 2000). The research and development outlined in this paper suggests several pathways towards a fundamental understanding of soundscapes with an emphasis on urban noise pollution and its effect on our cities. We discussed current and future system outputs that are designed to be scalable, transferrable, adaptable to other urban environments, and amenable to multidisciplinary engagement. This research represents a continued long-term effort of faculty, staff, and students from a number of organizations including the NYU Steinhardt, CUSP, Polytechnic School of Engineering, ITP; the California Institute of the Arts; and the NYC Department of Environmental Protection.

¹⁴ A crime prevention theory that has seen substantial success in crime control in large cities including New York, Chicago, Los Angeles, Baltimore, Boston, Albuquerque, Massachusetts, and also Holland.
5. References


Maisonneuve, N., Stevens, M., & Steels, L. (2009). Measure and map noise pollution with your mobile phone. *DIY:: HCI-A Showcase of Methods, Communities and Values for Reuse and Customization.*


titlecitygram: One Year Later ... In International Computer Music Conference Proceedings.
Sensing Urban Soundscapes. In Workshop on Mining Urban Data.
Environmental Health Perspectives, 108(Suppl 1), 123.
Peeters, G. (2004). {A large set of audio features for sound description (similarity and
classification) in the CUIDADO project}.
for spatial air pollution data. Environmental Health Perspectives, 111(10), 1283–8.
Selection of Mobile Sensors for Air Quality Monitoring.
320). Inner Traditions/Bear. Retrieved from http://books.google.com/books/about/The_Soundscape.html?id=_N56QgAACAAJ&pgis=1
http://scholar.google.com/scholar?q="let’s+try+to+keep+the+peace”&btnG=&hl=en&as_sdt=0,33#0
So-In, C., Weeramongkonlert, N., Phaudphut, C., Waikham, B., Khunboa, C., & Jaikaeo, C.
(2012). Android OS mobile monitoring systems using an efficient transmission technique
over Tmote sky WSNs. In Proceedings of the 2012 8th International Symposium on
doi:10.1109/CSNDSP.2012.6292763
for soundscape transformations. In Proceedings of GIS Ostrava 2013: Geoinformatics for
City Transformations. Ostrava, Czech Republic.
Recognition. In 2007 IEEE International Conference on Acoustics, Speech and Signal
doi:10.1109/ICASSP.2007.367191
considering their pass-by acoustic signature. The Journal of the Acoustical Society of
America, 133(5), 3322. doi:10.1121/1.4805552


“Big Data, Governance, and Equity”

Mai Thi Nguyen, Ph.D
Associate Professor
Department of City & Regional Planning
University of North Carolina at Chapel Hill
mai@unc.edu

and

Emma Boundy, M.A.
Department of City & Regional Planning
University of North Carolina at Chapel Hill

Prepared for

Workshops on Big Data and Urban Informatics
August 11-12, 2014

University of Illinois at Chicago
Chicago, Illinois

Please do not cite without permission. Please contact Mai Thi Nguyen for additional information: mai@unc.edu
Introduction

Big data is dramatically changing city hall. Elected officials and bureaucrats argue that big data allows for more efficient and effective decision-making. This has sparked a rise of Offices of Innovation, that collect, manage, use and share big data, in major cities throughout the U.S. A watershed moment for big data and cities was the announcement of President Obama’s Open Government Initiative in a memorandum in January 2009. The memorandum provided a Federal directive to establish deadlines for action relating to open data (Orzag, 2009). Thereafter, a rapidly growing number of local municipalities in the U.S. have engaged in activities related to making data more accessible (e.g. open data portals), developing policies around open data, and making government services and civic engagement easier through the use of new technologies and big data.

San Francisco launched the first open data portal for a U.S. city in the United States in 2009 and opened the first Mayors Office of Civic Innovation in January 2012 (Appallicious, 2014). As of July 2013, at least ten cities had Chief Innovation Officers, and a survey conducted in Spring 2013 found that “44 percent of cities of populations of more than 300,000 and 10 percent cities of populations between 50,000 and 100,000 had offices of innovation” (Burstein, 2013). While San Francisco had the first office, Mayor Bloomberg’s dedication to opening data in New York has been heralded by civic innovators as one of the driving forces behind the open data movement and the national trend towards greater civic entrepreneurship (Appallicious, 2014).

Offices of innovation have become popular because they offer the promise of using big data for predictive analytics, streamlining local government processes, and reducing costs. Yet, very little research has been conducted on what data is being harnessed, how it
is organized and managed, who has access, and how its use affects residents. Even less attention has been paid to the relationship between big data and equity.

This paper seeks to answer two questions. First, how is big data changing decision-making in city hall? Second, how is big data being used to address issues related to social equity? This paper seeks to answer these questions by examining offices of innovation that use big data in five major American cities: New York, Chicago, Boston, Philadelphia, and Louisville. In particular, this study examines three dimensions of big data and social equity: data democratization, digital access and literacy, and promoting equitable outcomes. Furthermore, this study highlights innovative practices that address social problems in order to provide directions for future research and practice on the topic of big data and equitable outcomes.

**Big Data, Governance, and Social Equity**

Although the private sector has become highly sophisticated at culling big data to shape business practices and planning for some time now, the use of big data in the public sector is a relatively new phenomenon. Much of the academic literature to date on big data and cities has largely focused on the historical evolution of big data and smart cities (Batty, 2012; Batty, 2013; Kitchin, 2014; Batty et al., 2012; Chourabi et al., 2012), the potential impact that big data can have on the future of citizens’ lives (Domingo et al., 2013; Batty, 2013; Chen & Zhang, 2014; Wigan & Clarke, 2013; Hemerly, 2013), and the challenges the public sector faces integrating big data into existing processes and strategies (Batty, 2012; Joseph & Johnson, 2013; Vilajosana et al., 2013; Almirall et al., 2014; Chen & Zhang, 2014; Cumbley & Church, 2013; Wigan & Clarke, 2013; Kim et al., 2014; Hemerly, 2013).
However, less research has centered on the relationship between big data, local governance, and social equity.

Social equity in governance, is defined by the Standing Panel on Social Equity in Governance as “The fair, just and equitable management of all institutions serving the public directly or by contract; the fair, just and equitable distribution of public services and implementation of public policy; and the commitment to promote fairness, justice, and equity in the formation of public policy” (National Academy of Public Administration, www.napawash.org). This definition of social equity focuses on the governance process and does not address equitable outcomes. For this study, we consider another dimension of social equity: public policies and government actions that promote greater equitable outcomes for the public. When social equity is tied to big data, a number of key themes emerge in the literature: digital access, digital literacy, and the use of big data to promote more equitable outcomes for the public. Thus, in our study, our definition of social equity considers both processes and outcomes.

Inequities in Access: The Digital Divide

As cities steadily transition towards a digital governance system, uneven access to digital technology across different groups may exacerbate social inequality. During the Clinton-era, the National Telecommunications and Information Administration (NTIA) raised awareness about the increasing “digital divide,” referring to the growing disparity in access to information technology (Light, 2001; Gilbert et al., 2008). Existing research on digital access has focused on how and why certain demographic groups have historically been left out of the technological adoption process (Batty et al., 2012; Chourabi et. al, 2012; Prieger & Hu, 2008; Gilbert et al., 2008; Lee and Hwang, 2014). Characteristics that can
significantly influence technological access and adoption are: age, income, race, education, and geographic location (Azari & Pick, 2005; Gilbert et al., 2008; Mossberger, Tolbert & Hamilton, 2012; Hilbert, 2011; Prieger & Hu, 2008; Revenaugh, 2000; Solomon, 2002; Velaga et al., 2012; Zhao, 2009). According to this research, the groups that have the least access to digital technology include: the poor, unemployed individuals, those with low levels of education, families without children, the elderly, non-whites, particularly non-Hispanic Blacks and individuals of Hispanic origin, and those living in rural areas (Dimaggio et al., 2004; Azari & Pick, 2005; Gilbert et al., 2008; Gilmour, 2007; Hilbert, 2011; LaRose et al., 2007; Lee and Hwang, 2014; Prieger & Hu, 2008; Prieger, 2013; Velaga et al., 2012).

While there is a large body of research on the digital divide, the focus of the research has been on broadband adoption, Internet usage, and computer access. To address this divide, federal funding has been used to develop initiatives to provide better access to computing centers and expanding availability of Internet services (Bailey & Ngwenyama, 2011; Revenaugh, 2000). Furthermore, local governments have attempted to improve computer and internet access for underserved population groups through partnerships with local schools and community centers and utilizing federal funding to improve local digital infrastructure in disconnected neighborhoods (Gilbert et al., 2008; Araque et al., 2013). Despite these efforts at eliminating the inequities in digital access, the digital divide still persists (Gilbert et al., 2008; Kvasny and Keil, 2006; Correa, 2010; Hargittai, 2002; Looker & Thiessen, 2003; Dimaggio et al., 2004; Light, 2001).
Digital Literacy: The “Participatory Gap”

While scholars have paid attention to the digital divide in relation to digital access over the past fifteen years, recent research has emphasized growing concerns over digital literacy, which can be considered the skills, knowledge, or familiarity with digital technology (Gilbert et al., 2008; Gilmour, 2007; Lee and Hwang, 2014; Correa, 2010; Hargittai, 2002). This digital divide has been referred to as the “participatory gap” (Fuentes-Bautista, 2013) and signifies that even if individuals have access to computers, smartphones, or the Internet, they may lack the skills, education, or familiarity to take advantage of the opportunities that information and communications technologies (ICTs) can provide (Warren, 2007; Gilbert et al., 2008; Kvasny & Keil, 2006; Looker & Thiessen, 2003; Dimaggio et al., 2004; Light, 2001). Differences in levels of accessibility and digital literacy are found to be correlated with typical patterns social exclusion in society (Warren, 2007; Lee et al., 2014; Mossberger et al., 2006; Dimaggio et al., 2004). In particular, socioeconomic status is considered the leading cause of the new literacy digital divide (Guillen & Suarez, 2005).

As municipal governments become increasingly reliant on digital technology, individuals that are not digitally literate will be unable to access municipal resources and services. The ability to navigate public agency websites, download and submit forms electronically, scan documents, and a host of other digital skills are increasingly becoming the default method of government operations, replacing paper and pencil. Lacking these skills will undoubtedly limit the ability of individuals to access resources and opportunities. Digital literacy not only affects individuals but local small businesses and organizations as well. Community-based organizations, for example, often struggle with having low capacity
to perform sophisticated studies or evaluations using big data and, therefore, do not have the ability to provide quantitative analyses that may be required to apply for and receive federal funding. Thus, understanding the barriers to digital literacy and the characteristics of groups and organizations that are persistently illiterate will allow local governments to adopt policies and practices to address it.

What do we know about the digital equity?

The first generation of initiatives developed to address the digital divide proposed that by improving digital accessibility, this would benefit disadvantaged groups and reduce gaps in access and usage (Azari & Pick, 2005). The idea behind these initiatives relied on the assumption that closing gaps in technological access would mitigate broader inequalities, including literacy. These initiatives assumed that by providing access to ICTs, this would improve disadvantaged groups’ social statuses (e.g. income) (Light, 2001). However, studies found other confounding factors associated with digital inequality: available equipment, autonomy in using ICTs, skills, support (e.g. technical assistance), and variations in purposes (e.g. using ICTS to obtain jobs vs. social networking) (Dimaggio et al., 2004). Increasing digital access does not adequately address these five issues and, therefore may be ineffective at reducing the digital divide (Looker & Thiessen, 2003). For example, Kvasny and Keil’s (2006) study found that providing computers, Internet access, and basic computer training was helpful, but not sufficient at eliminating the digital divide for low-income families in high-poverty areas. This study pointed to the intersection between digital inequities and other structural inequities in communities, such as lack of access to high-quality schools, limited public investment, and pervasive poverty. Even when digital divide initiatives do help low-income Americans living in poor neighborhoods
to gain digital literacy, these programs often do not mitigate inequities caused by disparities in transportation access or educational status (Light, 2001; Tapia et al., 2011). Thus, we believe that to be truly effective at reducing the digital divide, digital programs and policies must also be coordinated with other social policies that address the root causes of the digital divide: poverty, poor education, economic residential segregation, and public and private sector disinvestment in poor neighborhoods. Future analyses should consider how the potential of increasing access to and literacy of ICTs could make greater impacts on issues of equity if introduced in coordination with other innovative social policies and programs that provide extra support outside of the technological realm (Light, 2001).

Technological innovations have become an integral part of America’s communication, information, and educational culture over the past decade. Access to information and computer technologies is increasingly considered a “necessity” to participate in many daily functions (Light, 2001). As ICTs have become more integrated into daily life, populations that have historically not had access to or familiarity with how to utilize ICTs may become increasingly disadvantaged without improved access and literacy (Tapia et al., 2011). Furthermore, the disparities between those with and without digital access and skills can potentially have lasting, negative social and economic impacts on neighborhoods and cities.

The lack of access to computers, smartphones, and the Internet, as well as the related information that is accessible from these platforms, can exacerbate other forms of social, economic, and political marginalization for excluded groups (Gilbert et al., 2008; Lee and Hwang, 2014). The challenges of accessibility, usage, and low digital literacy for
various demographic groups will become increasingly important as cities innovate digitally, such as when cities integrate big data into existing processes (Scholl et al., 2009). For example, the potential of crowdsourcing as a citizen engagement platform will disproportionately benefit individuals and groups that participate in providing data through mobile applications or web-based applications. Without an understanding of how to use ICTs, already disadvantaged groups will not have their voices heard (Jenkins et al., 2009; Bailey & Ngwenyama, 2011; Lee and Hwang, 2014). However, being digitally illiterate may be of greater concern for more common procedures, such as job applications or qualifying for federal assistance programs. Today, even some minimum wage jobs now require job applications to be filed online. Individuals without access to a computer or the Internet and/or individuals without any familiarity in completing paperwork or forms online may experience significant difficulty completing the application, which may further exacerbate existing economic inequities. Data and digital inequities, in terms of both access and literacy, compound issues disadvantaged populations face. As a result, it is important to continually address inequity in digital access and literacy in order to prevent populations from becoming increasingly disenfranchised as technological innovations continue to develop.

Case Study Cities: Boston, Chicago, Louisville, New York and Boston

To better understand how big data is changing decision-making in city hall and how big data is being used to address issues related to social equity, we conduct interviews with key informants in five cities that have well-established offices of urban innovation: Boston, Chicago, Louisville, New York and Philadelphia. For more information about each city’s
office of urban innovations, see Appendix A. We used a snowball sampling design, which started with contacting the directors of each of the offices and receiving referrals from individuals we made contact with. We conducted a total of 19 semi-structured phone interviews with staff of the offices of innovations and key local stakeholders, such as users of big data disseminated by local governments, developers of new mobile applications that collect data, and staff of non-profits. These interviews lasted between 30 minutes to 1 hour and were transcribed. We supplemented our interviews with reports, newspaper articles, and scholarly publications that provided information about the five offices of innovations’ mission, goals, organizational and institutional structure, operational priorities, portfolio of programs, and key initiatives.

Offices of innovation are typically responsible for informational technology activities, such as providing Wi-Fi and broadband infrastructure, developing information technology policies, strategies, and benchmarks, and providing access to open data and mapping data through online portals. However, each of these offices has unique initiatives that have made them leaders in this field nationwide.

**How is Big Data Changing City Hall?**

**Big Data and Government Decision-Making**

The primary ways in which big data is changing City Halls nationwide are through the increasing the number of data sources (e.g. administrative, mobile application data, social media) to develop data analytics and predictive processes to inform decision-making. These types of systems are being developed to save money, allowing municipal governments to stretch budgets, improve efficiency, and develop new methods of
communication and networking internally in order to be more innovative in delivering public services. Two of our case study cities offer insights into how big data is used to inform local government decision-making: New York and Chicago. According to Stephen Goldsmith, Director of the Innovations in Government Program at Harvard, New York has been approaching big data from a problem solving approach, whereas Chicago is working to infuse data analytics into their existing governmental structure in a comprehensive way (Marks, 2013).

**Big Data and Predictive Processes: New York and Chicago**

New York City’s Mayor’s Office of Data Analytics (MODA) has been widely recognized for using predictive analytics in government decision-making over the past several years. Since 2012, New York has approached predictive analytics as a way to “evolve government” to improve the efficient allocation of resources and develop a better response to the real-time needs of citizens (Howard, 2011). New York City’s repository for administrative data is called DataBridge and was designed to perform cross-agency data analysis utilizing data from 45 city agencies simultaneously. According to Nicholas O’Brien, the Acting Chief Analytics Officer of New York, the main challenge the office has had to overcome has been working with big data from “45 mayoral agencies spread out in a distributed system...[this has been a challenging and arduous process because] each city department has a different anthology for how they characterize the data, so it’s important to understand the overlaps and the exceptions” (Personal Communication, February 24, 2014). Thus matching the data across administrative units and ensuring the quality of the data is extremely important as the decisions made through the predictive analytics process are only as valid and reliable as the data utilized to predict the outcomes.
Some of MODA’s most lauded successes include: “(1) a five-fold return on the time that building inspectors spend on looking for illegal apartments, (2) an increase in the rate of detection for dangerous buildings prone to fire-related issues, (3) more than a doubling of the hit rate for discovering stores selling bootlegged cigarettes, and (4) a five-fold increase in the detection of business licenses being flipped” (Howard, 2012). MODA’s impressive, quantifiable successes using predictive analytics has inspired other cities nationwide to create these types of processes to improve their own internal productivity and decision-making. Using data to inform these processes under an outcome-driven approach can help cities become economically and socially vibrant because the benefits garnered from this process are also helping to improve existing systems cities are already using.

In January 2014, Chicago received a $1 million grant from Bloomberg Philanthropies to create the first open-source, predictive analytics platform, called Smart Cities, that other cities will be able to utilize when it is complete (Ash Center Mayors Challenge Research Team, 2014). Chicago collects 7 million rows of data every day. This data is automatically populated and gathered in varying formats, through separate systems. The SmartData platform will be able to analyze millions of lines of data in real time to improve Chicago’s predictive processes. According to Brenna Berman, this platform will “develop a new method of data-driven decision making that can change how cities across the country operate” (Ash Center Mayors Challenge Research Team, 2014). The platform will perform two primary functions: (1) help city managers analyze trend data and conduct predictive analytics and (2) build capacity in cities that can not develop the platform on their own.
The SmartData platform will have the power to transform predictive analytics for cities nationwide through its open-source technology. If successful, the development of this replicable model for predictive processes can potentially change decision-making processes for every municipal government nationwide that chooses to utilize this software. The platform is being designed to be understandable to government employees with varying levels of data familiarity. According to Brenna Berman, “…we need to find a way to make analytics become available to the non-data engineer” (Shueh, “3 Reasons Chicago’s Analytics Could be Coming to Your City”, 2014). While Chicago has a team of data engineers, most cities do not have access to those resources and city staffers must make decisions without extensive experience in ICTs or data analytics. Developing a user-friendly platform that utilizes big data to develop predictive processes and analytics has the potential to reshape government decision-making for cities of all scales.

As previously mentioned, the private sector has typically been the primary user of big data, particularly for analytical purposes. IBM has ventured into the municipal analytics field by offering predictive analytics in the areas of traffic management, water management, emergency response, and building energy consumption (Batty et al., 2012; Shueh, “Big Data Could Bring Governments Big Benefits”, 2014). Each of these systems are tracked and catalogued in city governments nationwide, but establishing a platform to analyze this data in real-time and predict future patterns and behaviors requires increased capacity and skills, such as those provided by IBM, as well as an organization culture change. Strong-willed, data-driven leaders in city government are needed to undertake these types of major cultural changes and implement these transformative practices into existing data management policies (Shueh, “Big Data Could Bring Governments Big
Benefits”, 2014). According to Brett Goldstein, former Chief Data Officer of Chicago, predictive analytics is “having governments think about, ‘How do we prevent rather than react?’...Part of my role now is to say, ‘How can we use those techniques to do government better and do it smarter?’ (Rich, 2012).

*Using Big Data for Predictive Policing*

One area of predictive analytics that appears to adopted by cities is predictive policing, which employs data analytics to assist police make decisions that can prevent crimes, such as homicides, burglaries, and vehicle thefts (Novotny, 2014). Predictive policing systems use software that builds models that similar to private sector models of forecasting consumer behavior. Municipal police agencies use this technology to predict and prevent crime. The concept behind predictive policing is that situational awareness can be improved to create strategies that improve public safety and utilize police resources more efficiently and effectively (Hollywood et al., 2012). Employing limited resources more effectively and working proactively can help police departments anticipate human behavior and identify and develop strategies to prevent criminal activity. However, in order for predictive policing to be successful, the effectiveness of these outcomes must be quantifiable in order to specifically understand the impacts that data can make on these types of initiatives (Hollywood et al., 2012). Without quantifiable metrics, it will be more difficult to ascertain if changes in crime or public safety can be attributed to the predictive software or to another factor.

Two of our case study cities, Philadelphia and Boston, have already seen some quantifiable successes utilizing predictive policing. Philadelphia has been training police officers as data scientists in a “smart policing” program since April 2012. Officers complete
a two-week crime science program focused on utilizing technology to map crimes, understanding predictive software, and generating digital surveys to collect information from residents, which is a form of data crowdsourcing (Reyes, 2014). This program changes traditional police protocol because police officers, rather than external consultants, are directly trained in these technologies and build upon their skills and knowledge as a result of the program. “Part 1 crimes” (which include violent and property crimes), decreased by 5.8 percent and residential burglaries decreased by 39 percent in one district between 2012 and 2013. Philadelphia’s Deputy Commissioner, Nola Joyce, believes that the reductions in Part 1 crimes are due to the smart policing program and as a result of developing this program, the police department is moving from “counting and reporting crime” to “understanding” crime (Reyes, 2014).

Boston has established the Problem Properties Task Force, an interdepartmental effort to identify properties with persistent criminal activity and/or blight, that have caused problems (Boston’s Mayoral Transition: The Problem Properties Task Force, 2013). Developed to improve the allocation of the city’s limited resources, the Task Force convenes executive staff members from more than twelve departments and uses a data-driven, predictive analytics system that combines data from multiple city agencies into one database. As a result of this program, property assessment times have been shortened from days or weeks to seconds (Boston’s Mayoral Transition: The Problem Properties Task Force, 2013). The Problem Properties Task Force is notable because it is an example of multiple sources of city administrative data (big data) grounded in local knowledge from execute departmental staff to conduct predictive analytics and inform decision-making.
Balancing Predictive Analytics with Contextual Realities

While big data and predictive analytics offers the potential for greater efficiency and cost-savings, it can also do harm. Users of big data should be careful to ensure the accuracy and completeness of the data used in predictive models. There is also the potential for human error or misinterpretation of the results, thus it is important to cross check the findings from predictive models with individuals in the field – including staff who work within communities or the public at large. While efficiency is important, accuracy and transparency is equally important when using big data for predictive modeling or forecasting.

Big Data and Social Equity

The focus of predictive analytics in our case study cities has been on tame problems, such as infrastructure improvements, how to allocate staff time, and making city hall run more efficiently and proactively, rather than focusing on the more intractable problems of inequality, poverty and social equity. Based on our research, we find that there are three ways that cities are addressing social equity with big data: democratizing data, improving digital access and literacy, and promoting equitable outcomes using big data. We discuss each of these topics in turn below.

Data Democratization

Data democratization is centered upon the idea of increasing access to typically inaccessible or unpublished data for widespread analysis and consumption. There are many legislative forms in which data democratization can be encouraged or required by municipal governments. According to the Sunlight Foundation, of the 32 cities with open
data policies in place by April 2014, two are administrative memos, ten are executive orders, and 20 are “codified in legislation.” Codified in legislation is the “strongest” policy form, because “it preserves consistent criteria and implementation steps for opening government data beyond the current administration” (Williams, 2014). If an open data law is incorporated into legislation, consistency, enforcement, and management standards become part of the city’s legislation, which will be more difficult to overturn and alter when new administrations are elected. Regardless of the type of policy passed, each city’s relationship with open data is dependent on the municipal government’s structure and support for transparency, as well as the city’s existing mechanisms and capacity for data tracking and management.

*Open Data in New York City*

Of the five cities in this study, Philadelphia, Louisville, Boston, and Chicago have issued executive orders for open data. New York City codified open data into law. Enacted in March 2012, New York City’s landmark Open Data law--Local Law 11--was the first of its kind at the local U.S. municipal level (NYC DoITT, 2012). The result of Local Law 11 was that New York City established a plan with yearly milestones to release all of the city's data from city agencies by 2018. When finished, New York City will become the first U.S. local municipality with a complete comprehensive public agency dataset inventory (Williams, “NYC’s Plan to Release All-ish of their data,” 2013). According to Gale Brewer, the current Manhattan Borough President, as the first legislative open data effort, New York City’s law was more significant and transformative than the federal directive because it demonstrated how this type of work can be implemented at the local level (Goodyear, 2013).
The Mayors office of Data Analytics operates New York City's open data portal and works closely with NYC DoITT to populate the data portal and pursue other projects relating to data innovation and analytics (Feuer, 2013). MODA has been successful at procuring data from 45 out of over 125 public agencies thus far. Although New York City is at the forefront of making public agency data open, the process is still in the infancy stage. There are many challenges to completing this task, including the cost, organizational capacity, data management skills, and ongoing maintenance and upkeep of the data. What is also not clear is who will use the data and for what purpose. This raises questions about the format the data should be made available and the education that may be required to actually make the data valuable to a wide range of users. Nicholas O’Brien, Acting Director of MODA explains,

We’re also really starting to understand our audience. The customers of our open data portal are primarily non-profits, who we considered mid-tier data users that have some digital and data expertise but aren’t necessarily writing code or programming. We also know for our tech-savvy users, we have to direct them to our developer portal for more robust resources, and we have a third level of users that have very limited skills with data analysis. Understanding what each of these audiences want and need is an ongoing process for us.” (Personal Communication, February 24, 2014).

Open Data in Boston, Chicago, Louisville, and Philadelphia

Since 2012, Boston, Chicago, Louisville, and Philadelphia have established open data executive orders. The main forms of increasing access to data for these cities have been the development of open data portals and the creation of new executive positions to manage data initiatives. It is important to note that Philadelphia is the only municipal government in the country that does not “unilaterally control” the city’s open data portal (Wink, “What Happens to OpenDataPhilly Now?,” 2013). Instead, Philadelphia’s Open Data Portal is
managed by a Philadelphia-based non-profit and contains both municipal and non-municipal data (that users can submit directly).

In December 2012, Mayor Emanuel in Chicago established an open data executive order and created a position for a Chief Data Officer (CDO) to speed up the development of an open data portal. In order to improve transparency and build working relationships between departments with regard to big data, the executive order required that an Open Data Advisory Group, which includes representatives from each agency, to convene in order to discuss the portal’s ongoing development (Thornton, “How open data is transforming Chicago”, 2013). According to Brenna Berman, Commissioner and Chief Information Officer, meeting participants prioritize what datasets should be developed and identifies cross agency collaborations for data analytics” (Personal Communication, March 21, 2014). To support Chicago’s open data portal, the city established an accompanying data dictionary for information about the data being published. The Chicago Data Dictionary provides the public with an organized metadata repository that will eventually include metadata for every dataset published through the Open Data Portal (Thornton, “How Chicago's Data Dictionary is Enhancing Open Government”, 2013). The Data Dictionary takes transparency to another level and enhances the open data experience beyond what the other major American cities are doing. For interested citizens and developers that may be utilizing the municipal data, understanding the metadata for each dataset can be extremely beneficial.

In October 2013, Louisville announced an executive order for creating an open data plan. At that time, Louisville’s open data policy was the first U.S. municipal policy that stated open data will be the “default mode” for how government electronic information will
be formatted, stored, and made available to the public (Williams, “New Louisville’s Open Data Policy Insists Open by Default is the Future”, 2013). The implications are that data that is legally accessible will be proactively disclosed online through the city’s open data portal. Since January 2014, Louisville’s open data portal has been in development and operated by a small team working within the Louisville Metro Technology Services department (Holly Kessinger, Sharon Meador, and Kevin Landgrave, personal communications, February 25, 2014).

Among our case study cities, Louisville has the lowest population with almost 600,000 residents and the smallest city government. Currently, the city has a “homegrown portal” that the city staff developed. The current process for this homegrown portal requires Kevin Landgrave, Louisville’s data specialist, to determine (with a small team) which datasets should be prioritized based on how many citizens have been requesting them and how easily it will be to “clean” the data to become public. Louisville hopes to eventually publish between 500 to 1,000 datasets in the portal (Kevin Landgrave, personal communication, February 25, 2014). Unlike New York City’s law that mandates that all city agencies make their data accessible, Louisville relies on one staff member to collect, store, and manage the data. Due to this, the city must prioritize the type of data that is uploaded to the open data portal. Although the open data policy called for open data to be the “default mode,” they are far from developing the institutional infrastructure to accomplish this.

Data Democratization and Equity

Developing and maintaining an open data portal is a significant investment in terms of infrastructure and finances. One of the most common websites to manage municipal data
portals is Socrata. Boston, Chicago, and New York all use Socrata to operate their portals at a cost in the “high six-figures” for development and hosting the portal (Personal Communication, February 25, 2014). The city of Louisville does not have the funds to invest in a Socrata portal, which makes their “default” open data initiative a challenge to implement. As the trend towards open data continues, cities of all sizes are under pressure to adopt open data policies.

However, in the field of data democratization, opening data is only the first step. Data can be “opened” for broad impact, but cities need to learn how to leverage their data as a resource. According to Justin Holmes, the Interim Chief Information Officer in Boston, “we have to figure out how we take that data and make it more relevant. We know that Excel spreadsheets are not relevant to your grandmother. City departments need to be activists and understand why and how data can be impactful and then create a user-friendly platform” (March 19, 2014). While the open data movement has generated excitement and support from municipal governments, civic hackers, and tech-savvy citizens, these innovative applications typically provide benefits or services to those who also already utilize data and technology in their everyday lives. For citizens that have access to and understand these systems, they are able to receive benefits (in terms of cost, efficiency, and decision-making) that accessing data can provide.

Despite the publicity surrounding open data, providing data does not mean that every citizen will directly receive or experience a benefit or an improved quality of life. Truly innovative municipal governments should aim to create and engage a citizenry with widespread access and understanding of data and technologies (McAuley et al., 2011). By developing initiatives and services that benefit all population groups, as well as providing
access and training across a diversity of groups, cities can generate a sense of opportunity and improve the quality of life for its residents. This may require supplementary legislation to open more data relating to sensitive issues to inform decision-making in diverse fields, such as social work or housing.

The following analogy between libraries and open data portals is instructive for how data portals should be conceived: “we didn’t build libraries for an already literate citizenry. We built libraries to help citizens become literate. Today we build open data portals not because we have a data or public policy literate citizenry, we build them so citizens may become literate in data, visualization, and public policy” (Eaves, 2010). Nigel Jacob of Boston’s MONUM echoes these sentiments by saying “open data is a passive role for the government...fine for software development, but it does not actively engaged with citizens.” Thus, in order for cities to develop a democratic data system, they need to provide supplementary resources and training to ensure the widespread use and impact.

The New Digital Divide: Digital Literacy

In the last few decades, local governments have been engaged in activities to reduce the divide by increasing access to broadband, Wi-Fi, ICTs, and computer centers. Coupled with these programs and the rapidly lowering cost of acquiring technology, digital access is becoming less of a problem. As local government services and resources become increasingly digitized, it is important for residents to also be literate in new digital technology.

Thus, a new digital divide is emerging between individuals who are able to access and use digital resources and data effectively to improve their well-being and those who can not. For example, if an individual does not know how to download, fill out, and submit a
job application on-line, this will limit the types of jobs that they will apply for, thus limiting their opportunities. At the same time, opening vast amounts of data to the public is beneficial for tech-savvy users that understand how to manipulate and utilize that data for their personal or business gains. If digital literacy is low among groups that are traditionally disadvantaged, this may exacerbate social inequality.

*Chicago and New York: Digital Access and Literacy Initiatives*

In 2009, every city in our study, except Louisville received funding from the Broadband Technologies Opportunities Program (BTOP), a federal program designed to expand access to broadband services nationwide. New York City received $42 million from BTOP and developed the NYC Connected Communities program, which focused on improving public access to computer centers and broadband adoption in low-income and limited-English neighborhood throughout five boroughs (Personal Communication, March 10, 2014). Through this program, one hundred computing centers were opened at local public libraries, public housing, community centers and senior centers. The majority of these centers have remained open as more funding was acquired in 2013 when the BTOP funding expired. NYC Connected Communities included computer training and digital literacy programs that were specialized to the wants and needs of communities (Personal Communication, March 10, 2014). Since 2010, the NYC Connected Communities program has hosted more than 3 million user sessions citywide (NYC DoITT, “Technology & Public Service Innovation: Broadband Access”, n.d.). Approximately 100,000 residents have participated in training classes and over 4.7 million residents have attended open lab sessions (City of New York Public Computing Centers, 2014).
In Chicago, the Smart Communities program received $7 million of federal funding in 2010 to develop numerous training and outreach initiatives centralized in five low-income communities in Chicago (Tolbert et al., 2012). The Smart Communities program created a master plan that included considerable input from community members to determine program priorities, with each community creating their own projects to address challenges specific to their community (Deronne & Walek, 2010). Thus, in Chicago, the design of the programs was developed through a “bottom up” participatory process that resulted in unique programmatic components that focused on the idea that the “community knows best.” (Personal Communication, February 25, 2014).

Through early 2013, the Smart Communities program has trained approximately 20,000 people in computer and digital literacy skills (City of Chicago Public Computing Centers, 2014). One of the Smart Communities program’s most applauded successes is a statistically significant 15-percentage point increase in Internet usage in Smart Communities neighborhoods compared to other neighborhoods in the city between 2008 and 2011 (Tolbert et al., 2012). In Chicago, low broadband usage has consistently been correlated to income segregation, unemployment, recent school closing, and homicides (Knight Foundation, 2013), therefore the increase in internet use and improvement in computer and digital skills is benefitting some of the most disadvantaged neighborhoods in the city.

**Small-scale Approaches to Digital Access and Literacy**

The programs mentioned above are supported by large sums of federal funds. But, these funds are not available to the vast majority of cities throughout the country. Thus, we offer examples of smaller scale initiatives to improve digital access and literacy found in
our case study cities. In Chicago, LISC, a non-profit community-based organization, has built upon the Smart Communities program and developed community-focused initiatives that provide training for residents from a diversity of demographic backgrounds and offers an online presence for low-income neighborhoods. Boston and New York have installed computers in vans to serve as a mobile city hall. These vans are essentially mobile service centers that bring public staff into the field to offer services and to provide access to technology to residents of concentrated poor and minority neighborhoods. Boston operates “Tech Goes Home” (TGH), an initiative targeting school-age children and their families that provides digital literacy courses, subsidized computer software, and broadband access to participating families. Louisville is focusing on digital literacy from a workforce development perspective, as the city has made investments in developing high-level data analysis skills in residents that can improve employment possibilities for these residents while simultaneously attracting businesses to develop in Louisville.

The one issue with these smaller-scale approaches is that it is difficult to develop an initiative or program that tackles both digital access and digital literacy on a smaller scale and budget. The initiatives in Chicago have managed to continue this dual emphasis through LISC’s partnership with their other organizations. Boston’s Tech Goes Home program has managed to expand since its development in 2000 and has evolved into one of the more sustainable models of digital literacy by providing a computer and low-cost access to broadband to those who complete their program.

*Digital Access and Literacy Recommendations*

All the cities in this study have worked to close the digital divide in terms of access and literacy. However, the innovativeness and diversity of Boston and Chicago’s programs
demonstrate the significant investment and local resources required, both financially and in terms of the coordination between local stakeholders. Justin Holmes at Boston’s Office of Innovation & Technology’s describes the complexity of approaching issues of data access and literacy in diverse communities:

“Our engagement approach is multichannel...we need to be mobile, move beyond call centers and traditional centers, and use social media as a ‘value add’ to reach people. We're working to meet people where they are comfortable” (Personal Communication, March 19, 2014).

The main commonality between most of the cities was the development of public computing centers to improve access. Dan O’Neil, Executive Director of the Smart Chicago Collaborative, believes that “public computing centers are the most essential building block in providing access to technology” (Personal Communication, March 6, 2014). However, the programs with the potential for long-lasting impacts appear to be those with a concentrated effort on providing extensive on-site training through a site-specific curriculum tailored to the wants and needs of the community. Andrew Buss, the Director of Innovation Management in the Philadelphia Office of Innovation & Technology, identified that the key to Philadelphia’s KEYSPOT computing center initiative was having an instructor on site:

“You can't just have a room with a bunch of technology ...you need to have a person onsite at each location for assistance on how to use the equipment and to solve minor tech issues [which creates] a guided experience” (Personal Communication, February 28, 2014).

The availability and expertise of on-site instructors was also seen in each of the mobile van initiatives and has proven to be crucial for digital literacy programs. Furthermore, it seems that establishing a high level of trust between the program
providers, teachers, and participants is integral to the program's success and to see positive outcomes gained for students.

Improving data literacy is important for a diversity of users, not only user groups that do not have access to new technologies. Brenna Berman, the Chief Information Officer at Chicago's Department of Innovation & Technology, spoke about the importance of non-profits accessing and utilizing data:

“We’ve been creating a partnership between commercial organizations and the philanthropic community to make sure non-profits are benefiting from big data and using some indirect organizations that have been addressing the gap. The Chicago data ecosystem is aware of this challenge; we know non-profits were not embracing big data and weren’t using data to inform decisions. They needed representatives from communities to teach them how to do this, so we've run education workshops to collaborate and educate to make sure non-profits are in with this...like that saying, a rising tide raises all ships” (Personal Communication, March 21, 2014).

Therefore, closing the digital divide gap may not be simply a matter of providing access and training to individuals, but also to low capacity organizations.

Promoting Equitable Outcomes with Big Data

The third dimension of social equity relates to the promotion of equitable outcomes using big data. This could be conceived in two ways. First, directing big data analysis to reduce disparities across various social dimensions (i.e. income, race, ethnicity, and gender) for different groups. Second, targeting disadvantaged or underserved groups by using big data to improve their quality of life. The city of Louisville offers some innovative ways to address equitable outcomes using big data.

Sensor Technology in Louisville

Among government agencies, public health agencies have been leading the way in using big data to reduce health disparities. New technologies offer innovative ways to
assist low-income individuals to manage their healthcare and improve their health. In 2010, the city of Louisville created an inhaler sensor called Athmapolis, which also comes with a supplementary mobile application that was designed to improve asthma management for patients. Athmapolis allows asthma patients and their doctors to understand asthma’s triggers and provides an effective ways to control asthma, while simultaneously generating data for public health researchers (Propeller Health, 2013). More than 500 sensors have been deployed to low-income residents suffering with asthma in Louisville. While the program is still in the early stages, the benefits for residents have been notable. Interviews with program participants highlight their increased confidence in their disease management due to the “smart inhalers.” Furthermore, participants are happy to participate in the program because the inhaler sensor is provided free of charge through funding from philanthropic grants (Runyon, 2013).

Louisville is the only major city to implement this type of initiative on a cross section of their population and Ted Smith believes that these types of initiatives can be truly transformative on “breaking down data silos in the public sector...[which he believes] is a model project for how the public sector and communities should start working with informatics” (RWJF Public Health Blog, 2012). Utilizing this technology provides a benefit to the user, as their disease management improves. It also is useful to doctors and public health officials because individual-level data, geo-tagged by location is generated that can be utilized to inform future policy development. In Louisville, Ted Smith has pushed to incorporate innovations that improve public health because he believes having a healthy population contributes to regional and economic competitiveness, which encourages businesses to launch new offices in cities (RWJF Public Health Blog, 2012). This mindset
coincides with Louisville’s strategy of improving data literacy as a workforce development strategy, with the ultimate goal of improving the economic competitiveness of Louisville relative to other cities. Thus, for smaller cities such as Louisville, innovations in technology and big data that promotes the image and reputation of the city as cutting edge can improve economic competitiveness.

New York City: Improving Social Service Delivery

In 2009, HHS-Connect was developed in New York to collect all data relating to social services in one digital repository. The goal for this program has been to streamline the intake process for a client visiting different social service agencies and eliminate the need to reenter data and questionnaires. HHS Connect has transformed service delivery for social services into a client-centric model. The increased coordination between city agencies has improved coordination and delivery, increased accessibility to user data to improve case management processes, and provided clients with one access point to self-screen for over 30 benefit programs. These types of internal innovations can make the experience easier for clients while helping overburdened agencies detect fraud, improve service delivery, and reduce costs (Goldsmith, 2014).

These programs emphasize the need to develop partnerships between social service providers to allow data sharing between agencies to streamline services. In doing so, this not only creates organizational efficiencies but also makes receiving services for socially vulnerable populations easier and more efficient, thus saving individuals time and money. However, there are a variety of issues that limit the power of big data. On the federal level, statutes vary about what health records, educational transcripts, and data related to homelessness, child welfare, drug abuse and mental health can be collected, published or
shared (Goldsmith & Kingsley, 2013). On the state and local level, many laws were written prior to the digital age and can create conflict and confusion, thereby slowing down the adoption of innovations in these fields.

**Lessons Learned About Big Data, Governance, and Social Equity**

This study of five U.S. cities with offices of innovation sought to answer two primary research questions. First, how is big data changing decision-making in city hall? To varying degrees, big data in all of our case study cities is altering the way in which decisions are made in local government. The integration of data across city agencies is shown to create administrative efficiency and reduce man hours spent on tasks, thereby saving time, energy, and money. While this may be true, cities often do not talk about the costs associated with collecting, cleaning, managing, and updating big data. No study to date has examined the cost-effectiveness of these programs to determine what the return on investment actually is.

While big cities, such as New York, have high capacity public agencies that can populate the data required in a centralized repository, smaller cities may not. Louisville’s open data portal, for example, relies on one staff member populating the data using a data portal that was developed in house. What happens if this staff member leaves his post? New York City’s Mayor’s Office of Data Analytics provides the needed expertise and capacity to assist public agencies and departments to conduct predictive analytics. Their model of support is dependent on a separate government entity staffed with ICT experts dedicated solely to supporting other agencies. Other models of predictive analytics can be found in training an agencies staff member. Training police officers in predicting policing
analytics is an example of big data analytics altering the operations within a single public agency.

While there is great promise that predictive analytics and other types of programs to make local governments more efficient will be become widely accessible and affordable (e.g. Chicago’s Smart Data platform), caution should be taken to ensure that here are a checks and balancing when using big data. Big data analytics are only useful if the data is accurate and if the analysis of the data is context relevant. Therefore, big data analytics alone should not be used to make decisions, but rather, big data coupled with public engagement and experiential knowledge may make predictive analytics and decision-making more effective.

The second question is how is big data being used to address issues related to social equity? We examine three primary ways that big data relates to social equity. First, making data available and accessible can promote more social equity and open data portals are the primary way cities are doing so. New York City, having the only local open data law in the country, is at the forefront of this movement. Not only has the city institutionalized policies and practices around open data, but they have also facilitated it by adopting an internal database, DataBridge, and staffing the Mayor’s Office of Data Analytics with experts that can assist. New York City’s efforts with open data have been heralded by local and national experts as one of the most transformative and comprehensive among local municipalities.

While data democratization through open data portals have been a major focus for each city in this study, these portals do not necessarily promote equity. Open data portals provide data to the general public, but having data available does not mean that every
resident within a city is able to access, analyze, or use it for their benefit. Truly innovative municipal governments should aim to create and engage a citizenry with widespread access and understanding of data and technologies (McAuley, Rahemtulla, Goulding, & Souch, 2011). As described by Nigel Jacob of MONUM, “open data is a passive role for the government...fine for software development, but it does not actively engage with citizens” (personal communication, February 24, 2014). Under Nigel’s direction, MONUM does not operate Boston’s open data portal and instead focuses their resources on utilizing new technologies to improve citizen engagement experiences, education learning tools, and streetscapes.

None of the offices of innovation studied have programs that directly engage or teach the public how to maximize the potential of open data portals. The one organization that is attempting to teach marginalized populations about the benefits of open data is LISC Chicago. Since 2012, LISC Chicago has been offering “Data Fridays” at 3pm on Friday afternoons to teach interested users, of any digital literacy level, how to interact with data and technology and how these resources can be used to make neighborhoods stronger. Expanding these types of sessions through partnerships established between offices of innovation and local community groups would be one way of making the open data movement more equitable for all population groups.

The second dimension of equity, digital access and literacy, has been an area of concern for four of the five cities in our study. These cities received significant sums of federal funding and established initiatives aimed at improving digital access and literacy. However, upon the expiration of this initial funding, each city has struggled, to varying degrees, to continue these efforts. Chicago’s efforts at bridging the digital gap has continued
beyond the federal funding period due to the efforts of non-profits, such as LISC Chicago. Chicago’s Smart Communities Plan has established the framework for the municipal government to continue investing in broadband access and digital literacy programs, and LISC Chicago’s strong relationships with local community groups and residents helps with the implementation of this plan. Chicago’s work highlights the importance of federal funding, local planning, and effective collaborations with non-government organizations. The city of Boston also has innovative programs addressing digital access and literacy. Digital literacy courses such as Tech Goes Home is in high demand and has shown much promise. Furthermore, bringing city services to neighborhoods through City Hall to Go is reframing the relationship between Boston City Hall and the public by providing direct access to local municipal leaders and services on the van. Residents are able to interact directly with decision-makers, and benefit from spending less time and effort to receive municipal services when the van arrives in their neighborhood.

Understanding how big data can be used to address issues of equity is complex, due to the various dimensions of equity that can be considered. Each of the cities studied have been focusing on some issues of equity, but none of the cities have taken a comprehensive, multi-faceted approach to addressing numerous issues of equity. MONUM’s work in Boston primarily addresses social equity as their citizen engagement efforts target often overlooked population groups and their internal awareness of the need for improved engagement and service delivery for all citizens has the potential to shape their work moving forward. Another innovator of social equity is the SumAll Foundation in New York. While New York’s DoITT and MODA do not use equity as one of the main drivers of their work, the SumAll Foundation has found a way to utilize publicly accessible city data to
address the issue of eviction and shelter placements. Offices of innovation should examine SumAll’s approach to problem solving and look to combine their predictive processes capabilities to address issues that are directly impacting city residents.

Under the leadership of Ted Smith, Louisville has established groundbreaking initiatives using new technologies to address issues of environmental and health equity. Louisville’s emotional cartography and asthma mapping programs address several important components of innovating within municipal government. These projects were developed with support from other municipal government offices, local non-profits, local philanthropists, and partnerships with the University of Louisville. Utilizing the expertise, resources, and capabilities of each of these resources have made these initiatives possible, despite Louisville’s smaller size and budget constraints compared to the other larger cities studied. Louisville’s work also highlights another issue of data democratization. While the data gathered from these technologies can be used to inform health and environmental policy, the city realized that a new type of public-private portal needed to be established that can provide experts and policy makers with data from both public and private sources. Open data portals hosted by municipal cities typically only have the capability to host municipally-owned and generated data. Rather than be limited by these guidelines, Louisville is establishing separate websites and portals that can host datasets from multiple sources to address these issues of equity. As other cities begin to integrate the use of new technologies and multiple data sources to address issues of equity, the need for establishing partnerships between local stakeholders, as well as supportive digital entities that facilitate the organization and access to the big data generated, will become a necessity.
References


*NextBoston: Blog.* Retrieved April 30, 2014, from 
http://next.cityofboston.gov/post/65455635934/the-problem-properties-task-force-established-by

Retrieved April 6, 2014, from 
http://www.slate.com/articles/technology/future_tense/2013/06/big_ideas_for_cities_don_t_always_come_from_innovation_offices.html

Chen, C.L.P & Zhang, C.Y. (2014). Data-intensive applications, challenges, techniques and 
from http://dx.doi.org/10.1016/j.ins.2014.01.015

Chourabi, H., Nam, T., Walker, S., Gil-Garcia, J.R., Mellouli, S., Nahon, K., Pardo, T., & Scholl, 
of 45th Hawaii International Conference on System Sciences. Retrieved March 02, 

Telecommunications and Information Administration: Annual Performance 

Telecommunications and Information Administration: Annual Performance 

Correa, T., 2010. The participation divide among “online experts”: experience, skills and 
psychological factors as predictors of college students’ web content creation. J. 

http://dx.doi.org/10.1016/j.clsr.2013.07.007

Smart Communities. Retrieved April 9, 2014, from 
http://www.smartcommunitieschicago.org/news/2443

access to differentiated use. In K. Neckerman (Ed.), Social inequality (pp. 355–400). 
New York: Russell Sage Foundation.


March 8, 2014, from http://dx.doi.org/10.1016/j.telpol.2007.04.001


Appendix A: Case Study Cities

Boston Department of Innovation and Technology (DoIT)

The city of Boston has a unique structure for their office of innovation. Their office is called the Boston Department of Innovation and Technology or DoIT and is housed in Boston’s City Hall. Their primary role is to collect, manage, and organize big data. DoIT is also the city’s internal social media team and operates a coordinated, data-driven, strategy across all social media platforms, such as Facebook and Twitter, with the goal of curating daily engagement to help improve residents’ quality of life (“Boston’s Mayoral Transition”, NextBoston). The city has a social media policy and an organizational strategy to support this work with a social media liaison positioned in each of the city’s departments. Due to these efforts, Boston’s social media strategy has received national recognition and has seen exponential growth in terms of engagement with the public. For example, in 2013, the City of Boston’s Facebook page followers grew by 200% and the page’s reach grew 400% between 2012 and 2013 (“Boston’s Mayoral Transition”, NextBoston).

DoIT collaborates very closely with the Mayor’s Office for New Urban Mechanics or MONUM. According to the Co-Founder of MONUM, the department, “serves as a complementary force for city departments to innovative city services and we’re there to support them...[and unlike DoIT], MONUM has a great deal of independence and the ability to be innovative while not being encumbered by maintaining and supporting the innovation” (personal communication, March 19, 2014). Because MONUM is not managing the big data, the department focuses on piloting innovative, and sometimes risky programs that if successful, will be scaled up within a city department or city-wide. Thus, they are provided the freedom and flexibility to be creative and innovative. In 2013, Boston was named the #1 Digital City in America by the Center for Digital Government’s annual Digital Cities Survey (Department of Innovation & Technology, 2013). Between the efforts of DoIT and MONUM, and their social media strategy, the city of Boston is widely regarded as one of the leading big data innovators in municipal government.

Chicago Department of Innovation and Technology (DoIT)

In Chicago, the office of innovation is also known as the Department of Innovation and Technology (DoIT). This department takes a “comprehensive approach to data and analytics” and focuses their efforts of several key programs, including Chicago’s Digital Excellence Initiative, the Smart Communities program, and implementing Chicago’s Technology Plan, a comprehensive plan of five strategies and 28 initiatives to improve Chicago’s efforts in innovation and technology (Personal Communication, March 21, 2014). Since Brenna Berman was promoted to Chief Information Officer in late 2013, the department’s efforts have emphasized Ms. Berman’s personal vision of “resident-centered technology and innovation,” as well as internal innovations that foster data-driven decision making, such as predictive analytics programs, internal dashboards, and modernizing existing systems into user-friendly applications (Thorton, “Chicago Welcomes New CIO Brenna Berman”).
Louisville Department of Economic Growth and Innovation

In early 2012, Louisville’s Economic Development Department was restructured to become the Department of Economic Growth and Innovation. Ted Smith, previously the Director of Innovation, was appointed the Director of this new department. Louisville’s Department of Economic Growth and Innovation is a separate but coordinated department of the Louisville Metro Government, a regional government entity. According to Smith, the Department has three primary goals: (1) civic innovation, such as new approaches to community engagement, (2) government innovation, particularly innovating existing government processes, and (3) service of performance improvement, including creating internal dashboards, establishing cultural methodologies, and improving outcomes from a “bottom-up” perspective (Personal communication, February 27, 2014). Louisville’s Department takes a very unique approach to innovation, fusing economic growth and development principles and innovation. Much of the department’s efforts are focused on making the city more attractive to private businesses as well as developing digital platforms and infrastructure that can benefit both the city’s residents and the private sector (Raths, 2013).

New York Mayor’s Office of Data Analytics (MODA)

New York City’s innovation office is known as the Mayor’s Office of Data Analytics (MODA). MODA works in coordination with New York City’s Department of Information Technology & Communications (DoITT). NYC DoITT is primarily responsible for managing and improving the city government’s IT infrastructure and telecommunication services to enhance service delivery for New York’s residents and businesses. MODA was officially established by an executive order from Mayor Bloomberg in April 2013, but the agency had been working informally within New York City government for several years previously under the name “Financial Crime Task Force” (Personal Communication, February 24, 2014). MODA manages the city’s Open Data portal and works extensively on data management and analytics using their internally developed data platform, known as DataBridge. In order to establish DataBridge, MODA collaborated with DoITT to consolidate references for each building address in the city throughout all of the city’s agencies into one database, so that when one searches by address, all of the information from every department is accessible in one place (Nicholas O’Brien, personal communication, February 24, 2014). MODA operates specific projects to improve processes or gather more information about the city’s operations. MODA’s projects typically fall into one of these four categories: (1) aiding disaster response and recovery through improved information, (2) assisting NYC agencies with data analysis and delivery of their services, (3) using analytics to deliver insights for economic development, and (4) encouraging transparency of data between the city’s agencies, as well as to the general public (NYC Analytics, 2013).

Philadelphia Office of Innovation and Technology (OIT)

Previously known as the Division of Technology, Philadelphia’s Office of Innovation and Technology (OIT) was established in 2011 (Wink, “Office of Innovation and Technology to replace Division of Technology at City of Philadelphia”, 2013). Prior to reinventing and
restructuring the office to include innovation, the Division was responsible for the city’s
day-to-day technological operations. Philadelphia OIT was created as the city began
changing its culture and projects that were more innovative externally, as well as internally
within the infrastructure of municipal government. While Philadelphia OIT is responsible
for all major technology initiatives in the city, the department is broken into eleven units,
one of which is Innovation Management. The Innovation Management unit’s
responsibilities fall under three internal categories: (1) open data, (2) civic technology,
including mobile applications, and (3) innovation (Personal communication, February 28,
2014). The innovation category was established as the Philly KEYSPOT initiative was
launched, a federally funded public-private partnership that established approximately 80
public computing centers in communities and provides residents with Internet access and
training.

As mentioned earlier, the city of Boston’s Mayor’s Office of New Urban Mechanics
has a satellite office in Philadelphia with the same name. According to Almirall et al.
(2014), Philadelphia’s MONUM is referred to as a civic accelerator, which is an organization
that “match(es) cities with start-ups, private firms, and non-profit organizations interested
in partnering with government to provide better services, bring modern technology to
cities, or change the way citizens interact with city hall” (p. 4). MONUM’s Philadelphia
office is also located in city hall and its mission to transform city services and engage
citizens and institutions throughout the city to participate in addressing the needs of city
residents.

David A. King, Ph.D.
Assistant Professor of Urban Planning
Graduate School of Architecture, Planning and Preservation
400 Avery Hall
MC 0356
1172 Amsterdam Avenue
Columbia University
New York, NY 10027
dk2475@columbia.edu

Jonathan R. Peters, Ph.D. (corresponding author)
Professor of Finance
The College of Staten Island
and
Research Fellow,
University Transportation Research Center – Region II
The City College of New York

Room 3N-207
2800 Victory Blvd.
Staten Island, New York, 10314
Phone: 718-982-2958
Fax: 718-982-2965
Email: jonathan.peters@csi.cuny.edu

Adam Davidson, M.C.P.
Ph.D. Program in Earth and Environmental Science
The City University of New York
365 Fifth Avenue
New York, New York 10016

Juan Francisco Saldarriaga
Associate Research Scholar
Adjunct Assistant Professor of Urban Planning
Spatial Information Design Lab
GSAPP | Columbia University
(212) 854-2783
Abstract

Cities across the United States are developing and installing bike share systems to improve access to cycling. The largest system in the country is in New York City, with almost 6,000 bikes across more than 300 stations, but most systems are a few hundred cycles sprinkled throughout a limited number of docking stations. Though modern bike share represents a new transportation technology, these systems quickly became fixtures in the cities where they operate. Despite their popularity, these programs require municipal support, which may include direct financial support, allocation of street spaces for docking stations, or other types of direct and in-kind assistance. The public involvement raises concerns about how the systems are designed and who has access to the bikes. More succinctly, do publicly supported bike share systems adequately serve a city’s population?

We explore the distribution of bike share in three ways. First, we examine the spatial networks of bike share systems in New York City, Boston, Chicago, Washington, D.C. and California’s Bay Area. We use GINI coefficients to show that the socio-demographic characteristics in areas served by bike share networks are similar to those areas served by public rail transit networks. Second, we use demographic splits to show how New York’s Citibike program tilts heavily toward male users, but the fee structure is somewhat biased against women. Third, we proposed a credit-based system rebalancing program that both reduces rebalancing costs and improves gender equity. Overall this research demonstrates how new ("Big") data can be incorporated with more traditional form of analysis to both evaluate and improve the equitable distribution of publicly
available assets. We conclude with a discussion of the implications of this work for subsidy policy and local transportation planning.

Introduction

The recent success of bike share systems is part of a larger trend toward multimodal transportation systems in cities. City officials and planners focus on public health benefits of active transportation, the low cost of access relative to other modes of travel, the imperative to promote sustainable transportation policies, and the increasing size of bicycle lane networks[1, 2]. These trends help shift transport planning away from auto-orientation and toward transit, walking, cycling and other services.

However, there have been limited studies of effect of bikeshare programs on equity needs. A 2013 evaluation of Philadelphia’s 2010 Bikeshare Concept Study focused on the potential public health impacts of a system that still has not been launched [3]. The research noted high concentrations of asthma, obesity, heart disease, and other health concerns when viewed by resident geography, but there are only small overlaps of proposed bike share stations with neighborhoods with the poorest health. The author concludes that “the program’s planners must take additional steps to ensure that those who are most likely to develop the health conditions against which biking can offer protection have access to the program”[3].

Early evidence suggests that planning bike share with an eye toward public health has either not been taken or do not provide an appreciable result. A survey of Washington, D.C.’s Capital Bikeshare users showed that riders were more likely to be
young, white, male, well educated, and only slightly less affluent than the regional
average [4]. In a Masters thesis that focused on modeling the determinants of usage for
Capital Bikeshare, Daddio found a similar correlation by examining the demographic
makeup of a station’s catchment area relative to trip departures per station. His most
significant results suggest that nearly 80% of the variation in ridership can be explained
by five variables: the positive effects of being proximate to a young population (age 20 –
39), proximity to a Metro station and a retail area; and the negative effects of being
distant from the system center and proximate to a non-white population [5]. In short,
Downing’s conclusion is that the communities with the most to gain are the ones that
Daddio finds use the system least.

Other evidence suggests that bike share systems can act as complements to
conventional transit services and substitutes for some auto trips [6].

Data

Bike share systems in the United States use similar technologies, often from the
same suppliers, and generally are supportive of open data standards [7-11]. Overall, bike
share data used comes from two sources: historical data on individual trips and stations
provided through each company’s website and live feeds displaying individual station
data. Not all of the systems use both sources. While some of them update and offer their
data permanently, others, such as Divvy and Bay Area Bike Share have only made their
data available in the context of data visualization challenges open to the public. This
being said, the availability of the data is very good and as it always comes with
geographic coordinates for the bike share stations, it is very simple to place them on a
map using ArcGIS software.
Below we describe data used for each system.

- CitiBike (New York) provides both individual trips and a live feed updated every two minutes in json format. The location of the stations can be derived from both, although the live feed includes stations that might not be in use. With the individual trips file, you can use only the stations that have actually been active during that month. The data is for the period between July 2013 and May 2014.

- Divvy (Chicago) made their trip and station data for a data visualization challenge earlier this year. The data is for 2013 and it includes separate csv files for individual trips and stations and a GIS shapefile with the stations.

- Hubway (Boston) also made its data available in the context of a data visualization challenge in 2012. However, that dataset has been updated and it now includes individual trip data from October 2012 to November 2013. The dataset also includes a file with individual stations. All files come in csv format.

- Bay Area Bike Share (Bay Area) also made its data available in the context of a visualization challenge in early 2014. The dataset includes individual trip data from August 2013 to February 2014, as well as individual station data and rebalancing data, all in csv format.

- Capital Bikeshare (Washington D.C.) provides their individual trip data on their website
as well as an xml station status feed. The trip data includes individual rides from September 2010 through March 2014.
Overview of the Bikeshare Networks

In this section we focus on five systems currently operating in the United States. Figures 1-5 show the spatial distribution of docking stations for New York City, Chicago, the San Francisco Bay Area, Boston and Washington, D.C. Each city’s system shows distinct spatial characteristics of coverage, network density and connectivity. New York’s network of docking stations is very dense, where most stations are with a few blocks of another. Boston’s and Chicago’s systems feature stations much more spread out than New York’s, but both are contiguous. Washington and the Bay Area have spatially dispersed stations and feature spatial discontinuity where clusters of stations operate somewhat independently.

The network design of bike share systems has implications for usage, rebalancing issues, and equity. Nodes and stations within bike share networks are deliberately placed by urban planners and local officials. CitiBike in New York, in conjunction with the city’s Department of Transportation, encouraged the public to submit station requests through an online portal [12]. Yet despite the enthusiasm of the crowd across the city, a dense network centered on lower Manhattan and Brooklyn was implemented and most of the proposed station areas were left without. While most of the city is far away from the bikes, the distribution closely aligns with wealthier parts of the city. In areas of Brooklyn to the east of the shaded area on the map the bike share stations do not serve lower income neighborhoods. The deliberate spatial distribution raises concerns that the bike share system is a new alternative for wealthier New Yorkers. While such a statement in factually true—CitiBike is distributed in wealthier parts of the city—the policy implications are ambiguous.
CHICAGO - DIVVY

AREA WITHIN HALF A MILE OF DIVVY STATIONS

- DIVVY STATION

HALF A MILE RADIUS

0 2.5 5 Miles
BAY AREA - BIKE SHARE

AREA WITHIN HALF A MILE OF BAY AREA BIKE SHARE STATIONS

- BAY AREA BIKE SHARE STATION

HALF A MILE RADIUS

SAN FRANCISCO

REDWOOD CITY

PALO ALTO

MOUNTAIN VIEW

SAN JOSE
BOSTON - HUBWAY
AREA WITHIN HALF A MILE OF HUBWAY STATIONS
- HUBWAY STATION
- HALF A MILE RADIUS
Overall bikeshare networks are centered on central business districts rather than residential neighborhoods. However, in some instances the networks are more proximate to non-whites than the more extensive rail system. We use U.S. Census data form the American Community Survey and the Longitudinal Employer-Houshold Dynamic (LEHD) to examine the spatial relations between bike share, households and employment. To keep comparisons consistent, Boston is also left out of some summary
data. As complete metropolitan areas these regions (excluding Boston) contain more than 40.5 million people and nearly 17 million primary jobs.\footnote{The state of Massachusetts did not submit LEHD encoded data, therefore it cannot be included in employment tabulations.}

The strength of using the Census and LEHD data is that we can compare & contrast residential and worker populations, thus giving a more complete picture of who has potential geographic access to transportation infrastructure throughout the course of their day. However the datasets provide some general, but not exact comparisons.

Detailed census data is mostly available at the tract level and above, while LEHD data is based on the census block group of the employer. Census data is on the whole more complete, providing detailed income, age, and racial data. LEHD data uses broader bins to group income, age, and race than the census. For example LEHD income is grouped into counts of three categories: “Below $1,250 / month”, “Between $1,251 - $3,333 / month”, and “Above $3,333 / month.” Census data provides median income by tract, as well as counts at much finer and longer gradations.

With these caveats in mind the authors looked at background population and employment characteristics of areas within a half mile of the bikeshare systems and then compared that to areas within a half mile of metro or subway style rail transit. This comparison is based on the common claim that bikeshare serves as a complement to transit, since it helps with the ‘last mile’ issue, short trips and circulation. For these reasons, areas with only commuter rail were not included since it serves longer distances and nearly exclusively the work trip. Likewise, areas only near bus service were also not included as many bus lines are perceived to suffer from speed, frequency, and reliability and thus may not be perceived as complements or substitutes to bike or rail transit by
many potential riders. The resulting figures were then normalized as a percent of the total population category within the MSA.

Across selected regions (excluding Boston) rail is within a half mile of 18% of the residential population. However, if we also drop New York we bring that average to 11.5% (if we then add in Boston the number is only slightly raised to 12%). By comparison bikeshare is only accessible to about 6% of the residential population, but that ranges from a low of 3.6% of metropolitan San Franciscans, to a high of 13% of metropolitan Washingtonians. For Washington, that is actually greater coverage than the rail system provides. However, these numbers are low when compared to the number of jobs accessible to rail and bikeshare. Averaging the five regions where job numbers are available, we find that 33% of jobs are within a half mile of rail compared to 25% for bikeshare. In Washington and Chicago, the numbers are nearly even with bikeshare trailing rail by just a couple of points.

Table 1: 
*Does not have LEHD employment data, so is consequently left out of the Summary column

<table>
<thead>
<tr>
<th>AllPop (100%)</th>
<th>BOS*</th>
<th>CHI</th>
<th>DC</th>
<th>NYC</th>
<th>SF</th>
<th>SJO</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AllJobs (100%)</td>
<td>4,552,402</td>
<td>9,461,105</td>
<td>5,582,170</td>
<td>18,897,109</td>
<td>4,335,391</td>
<td>1,836,911</td>
<td>40,112,686</td>
</tr>
<tr>
<td>Pop_NonWhite</td>
<td>964,862</td>
<td>3,277,224</td>
<td>2,523,350</td>
<td>7,719,445</td>
<td>2,095,872</td>
<td>965,114</td>
<td>16,581,005</td>
</tr>
<tr>
<td>Pct_Pop_NonWhite</td>
<td>21%</td>
<td>35%</td>
<td>45%</td>
<td>41%</td>
<td>48%</td>
<td>53%</td>
<td>41%</td>
</tr>
<tr>
<td>Job_NonWhite</td>
<td>837,339</td>
<td>808,267</td>
<td>2,245,480</td>
<td>648,368</td>
<td>306,854</td>
<td>4,846,308</td>
<td></td>
</tr>
<tr>
<td>Pct_Job_NonWhite</td>
<td>21%</td>
<td>31%</td>
<td>29%</td>
<td>35%</td>
<td>38%</td>
<td>29%</td>
<td></td>
</tr>
<tr>
<td>Pop_Young</td>
<td>1,253,403</td>
<td>2,649,924</td>
<td>1,642,101</td>
<td>5,266,518</td>
<td>1,254,495</td>
<td>537,456</td>
<td>11,350,494</td>
</tr>
<tr>
<td>Pct_Pop_Young</td>
<td>28%</td>
<td>28%</td>
<td>29%</td>
<td>28%</td>
<td>29%</td>
<td>28%</td>
<td></td>
</tr>
<tr>
<td>Job_Young</td>
<td>879,612</td>
<td>604,973</td>
<td>1,672,195</td>
<td>365,635</td>
<td>152,054</td>
<td>3,674,469</td>
<td></td>
</tr>
<tr>
<td>Pct_Job_Young</td>
<td>22%</td>
<td>23%</td>
<td>22%</td>
<td>20%</td>
<td>19%</td>
<td>22%</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Size of Systems

<table>
<thead>
<tr>
<th>Metro</th>
<th>Rail</th>
<th>Bikeshare</th>
<th>Rail / Metro</th>
<th>Bikeshare / Metro</th>
<th>Bikeshare stations</th>
<th>Bikes</th>
<th>Docks</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOS</td>
<td>3487</td>
<td>43.1</td>
<td>25.5</td>
<td>0.012</td>
<td>0.591</td>
<td>131</td>
<td>1300</td>
</tr>
<tr>
<td>CHI</td>
<td>7197</td>
<td>72</td>
<td>44.1</td>
<td>0.01</td>
<td>0.613</td>
<td>300</td>
<td>3000</td>
</tr>
<tr>
<td>DC</td>
<td>5598</td>
<td>52.3</td>
<td>71.4</td>
<td>0.009</td>
<td>1.363</td>
<td>321</td>
<td>2500</td>
</tr>
<tr>
<td>NYC</td>
<td>6687</td>
<td>120</td>
<td>16.8</td>
<td>0.018</td>
<td>0.14</td>
<td>330</td>
<td>6000</td>
</tr>
<tr>
<td>SF</td>
<td>2471</td>
<td>43.4</td>
<td>6.2</td>
<td>0.018</td>
<td>0.144</td>
<td>42</td>
<td>700</td>
</tr>
<tr>
<td>SJO</td>
<td>2679</td>
<td>34.6</td>
<td>8.5</td>
<td>0.013</td>
<td>0.246</td>
<td>28</td>
<td>700</td>
</tr>
<tr>
<td>Summary</td>
<td>28119</td>
<td>365.4</td>
<td>172.5</td>
<td>0.013</td>
<td>0.472</td>
<td>1152</td>
<td>13500</td>
</tr>
</tbody>
</table>

Evaluating Equity with GINI Coefficients
Work by Gordon and Peters address the issue of incongruity of users. In examining survey data of tolled bridges in the New York City region, they find that facilities with alternatives (good mass transit or free nearby routes) show higher levels of income inequality than their surrounding area, while facilities with few alternatives better reflect the background demographics of the area that they are situated in [13]. Thus facilities that are either inherently accessible or have few alternatives will share the demographic characteristics of their surroundings. Applied to bikeshare, we would not expect bikeshare to reach lower income or underrepresented groups if it is not first made geographically accessible to them. This adds an additional tension to Daddio’s findings by indicating that there is a tension between maximizing ridership and social equity.

Gordon and Peters explored the use of geographic information systems as a tool to use for the evaluation of social equity. They estimated the relative burden of tolling on various user groups and estimated the general markets for transportation services. Using Lorenz curves and GINI coefficients, Gordon and Peters are able to estimate the income profiles of the general population that are geographically located near the toll facilities and the actual users of particular toll bridge facilities.

Likewise, the authors propose to estimate the likely users of bikeshare – that is, the population that is proximate to the bikeshare stations and following Daddio’s findings, we expect that proximity to the system and centrality to the geographic center of the bikeshare network represents the users who have the best opportunity to utilize the bikeshare system. The authors pulled the Census 2010 data from the existing bikeshare stations in the four study cities and estimated the relative income distribution of the bikeshare systems in New York City.
New York’s Citibike System has been the subject of considerable interest regarding the financial needs of the system. Given that the New York system is not subsidized in terms of direct system operational support, the financial need to create enough revenue to cover existing costs has been a matter of major concern. Recently, the system has reported an ongoing operational loss. One consideration is the ability to focus additional fees for service to address the need for revenue and consider options as to how to raise revenue from the various classes of users – while avoiding any negative impacts on the social equity of the system. As highlighted above, the system network have in some cases a more limited footprint in terms of low income users as compared to the general metropolitan area (Washington DC excepted).

Social Justice analysis is mandated by Executive Order 12898 of 1998 and by Title 6 of the Civil Rights Act of 1964. These federal requirements are intended to assure that disadvantages classes are not discriminated against in terms of various government services. In the transportation sector, these rules are general understood to require that agencies address social justice concerns in the planning and implementation of transportation systems. Given that bikeshare systems are deployed under municipal contract, utilize public road space for stations and in many cases receive public subsidies as stated above, it is clear that social justice analysis is appropriate and may well be mandated by federal authorities. As such, consideration of how a change in rates should be applied or what aspect of costs subsidized should be evaluated for social justice issues.

In particular, the key metrics of analysis of social equity analysis are as follows:

1) Analyze the needs of protected classes
2) Included in the transportation planning process
3) Protected classes are not overly burdened with the cost of services
4) Protected classes are not denied the benefit of transportation services

In terms of protected classes, one generally considers the following as protected classes:

1) Low income individuals
2) Minority individuals
3) Disabled individuals

<table>
<thead>
<tr>
<th>Location</th>
<th>Buffer</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC</td>
<td>Capital Bikeshare</td>
<td>0.4535</td>
</tr>
<tr>
<td>DC</td>
<td>Bus Stops</td>
<td>0.4326</td>
</tr>
<tr>
<td>DC</td>
<td>Metro Stations</td>
<td>0.4443</td>
</tr>
<tr>
<td>Chicago</td>
<td>Bikeshare</td>
<td>0.4949</td>
</tr>
<tr>
<td>Chicago</td>
<td>Bus Stops</td>
<td>0.4809</td>
</tr>
<tr>
<td>Chicago</td>
<td>Rail Stations</td>
<td>0.5020</td>
</tr>
<tr>
<td>Boston</td>
<td>Bikeshare</td>
<td>0.5028</td>
</tr>
<tr>
<td>Boston</td>
<td>MBTA</td>
<td>0.4965</td>
</tr>
</tbody>
</table>

Revenue Generation from Various Fees – Social Equity Conditions

<table>
<thead>
<tr>
<th>Location</th>
<th>Buffer</th>
<th>Gini</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC Capital Bikeshare</td>
<td>0.4535</td>
<td></td>
</tr>
<tr>
<td>DC Bus Stops</td>
<td>0.4326</td>
<td></td>
</tr>
<tr>
<td>DC Metro Stations</td>
<td>0.4443</td>
<td></td>
</tr>
<tr>
<td>Chicago Bikeshare</td>
<td>0.4949</td>
<td></td>
</tr>
<tr>
<td>Chicago Bus Stops</td>
<td>0.4809</td>
<td></td>
</tr>
<tr>
<td>Chicago Rail Stations</td>
<td>0.5020</td>
<td></td>
</tr>
<tr>
<td>Boston Bikeshare</td>
<td>0.5028</td>
<td></td>
</tr>
<tr>
<td>Boston MBTA</td>
<td>0.4965</td>
<td></td>
</tr>
</tbody>
</table>

Proc. NSF Workshop on Big Data and Urban Informatics, Nov. 2015 Page 957
4) Senior Citizens
5) Rural populations

In addition, the impact on gender equality needs to be considered as we discuss changing various aspects of the system. Given the existing demographic data collection on annual users (subscribers) as opposed to daily and weekly users (customers), we can evaluate the existing subscribers in terms of gender and age issues. The authors developed a detailed record by record pricing algorithm to price the individual rides on the Citibike system in terms of overtime fees. We also utilized the demographic information and reported subscriber rates by gender to understand the fee structure aspects of bikeshare use. We evaluated 5,561,840 rides from June 2013 to February 2014. In addition, we also explored the cost of annual membership fee.

New York’s bikeshare system has benefited from high usage and enrollment as compared to other systems. However, the data indicates that a large amount of rides for subscribers are provided at no additional cost to the users other than the annual fee. Table 2 provides an overview of the revenue and usage components by gender for the trips reported from the system data.

The data reports a considerable amount of variation in usage pattern and fee structure by gender. Women represent 38% of Citibike subscribers but only 23.2% of rides. This strong imbalance provides one with an argument that additional fees are better applied to a per ride costs (say a flat fee of $.25 per ride) as opposed to increases in the annual subscription fee. Further, additional subsidies are best applied to lowering the annual fee as opposed to general subsidy on all rides.

As a second consideration, one could consider increasing the overtime fee for rides that exceed the 45 minute basic ride limit. Examining the mix of overtime fee payments, we find that women pay 30.3% of the overtime fees – as they have a greater frequency of overtime rides with 1.17% of rides by women having an overtime fee as compared to .79% of rides by men. Given this disparity, increases in overtime fees also appears to have a significantly disparate impact on female users of this bikeshare system.

Taken as a whole, it appears due to variation in usage patterns that changes in the fee structure of the New York Citibike system should be evaluated for social justice impacts in terms of gender. If additional information was collected on racial and income data from subscribers one could further evaluate additional environmental justice issues.

Table 2 – New York City Bikeshare Data by Gender

<table>
<thead>
<tr>
<th></th>
<th>Subscriber</th>
<th>Rides</th>
<th>OT Fee</th>
<th>Rebalance Net</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subscribers - Male</td>
<td>62%</td>
<td>76.8%</td>
<td>69.7%</td>
<td>94%</td>
</tr>
<tr>
<td>Subscribers - Female</td>
<td>38%</td>
<td>23.2%</td>
<td>30.3%</td>
<td>6%</td>
</tr>
</tbody>
</table>
A further matter of considering additional fees follows as we examine the potential of using the pricing system to manage the distribution of bicycles in the system.

**System Rebalancing**

One aspect that can be informed by system performance metrics is process and operational improvements. Bikeshare systems, while deployed for a considerable amount of time in a number of cities, still has a number of areas of operation practice that could benefit from additional analysis and exploration of usage patterns. One area of interest is the problem of system rebalancing. It appears to be a rather common issue that bikeshare system can suffer from bicycle shortages and surplus conditions in various stations at different times during the day and on specific days of the week. If we examine these patterns of usage, we find that these patterns appear to be stable and occur repeatedly at key stations in the network.

These oversupply or shortage conditions potentially can restrict system usage and also can create difficulties for users who may want to check in equipment at overloaded stations. The current solution to these problems is to physically redeploy the bicycles from surplus stations to shortage stations. Yet information on this aspect of bikeshare systems is difficult to examine. It is known through anecdotal comments that systems apparently have these problems, but the reporting on the amount of rebalancing needed for a given system is generally not reported in existing data sources. Dantos (2012) produced some analysis that indicated a greater propensity for users to move downhill as opposed to uphill from Capital Bikeshare stations. His work indicates a greater outflow from stations in Northwest Washington (an uphill area) as compared to other regions. The net result of this is a need to physically redistribute the bikes as needed to address this asymmetrical demand and supply conditions. This is similar in practice to the need to deadhead buses or trains in a transit system to redeploy assets as needed to address system load condition.

The scale and magnitude of these redeployments may be the subject of internal consideration for bikeshare operators, but public discussion and policy aspects appear to be understudied. Clearly, the need for excessive manual movements by trucks of bikeshare equipment impact the carbon footprint of the system and the cost of operations. Yet some information is available and provides tantalizing insight into the rebalancing issue. Recently, Alta, the CitiBike contract provider was subject to a lawsuit for back wages for rebalancing drivers and mechanics. As reported in the Oregonean, one employee stated that "A large scale bike-share program doesn't work without a fleet of trucks to keep the inventory balanced,". Further analysis of this problem is clearly merited. In most systems, the additional bike movements caused by rebalancing is scrubbed out of the data that is reported for public use.

The authors examined this question utilizing the existing volume data and developed metrics of oversupply and shortage by station for the Citibike System. As reported by
others—such as Dantos—many stations had reasonable balance in terms of bike inflows and outflows. These stations required little rebalancing on a daily basis and as such incurred less operational costs for this aspect of the system. We examined all 330 stations in the CitiBike system and located a number of stations that had significant imbalances. It is interesting to note that while many of the imbalance hours occurred during the morning and afternoon peak, some imbalance conditions existing for extended periods of time. In addition, during the same period some stations exhibited surplus conditions and others shortage conditions. As such, these conditions appear to be linked to various user needs and the regional job and home relationships.

One method to deal with this structural imbalance is to alter the price of dropping off or checking out a bike based upon the general daily pattern of usage by station. In our first example, we apply a fee of $1.00 to take a bicycle out of a station during a period of shortage. Further, we apply a credit of $1.00 for checking a bicycle into the same station during the same period—thereby increasing the supply of bikes at a deficit station. Correspondingly, we apply a credit of $1.00 to user accounts if they check a bike out of a station during a surplus period and a fee of $1.00 for checking in a bike at a surplus station during surplus demand hours.

The authors developed a pricing algorithm that reviewed the trip records and applied the fee or credit to three test stations which have very strong imbalance problems. The three stations have varying patterns of demand, with one having a surplus condition in the AM Peak (6:00 to 10:00 AM) and deficits conditions in the afternoon/evening. The other two stations have deficits in the AM peak and surplus conditions in the afternoon/evening. There were 140,539 trips that had a fee or credit applied producing 44,830 dollars in credits and 87,495 in fees for a net subscriber cost of $42,665 in net revenue over 7 months. This revenue was generated in a very asymmetrical way in terms of gender, with 94% of the net revenue generated from male subscribers and 6% from female subscribers.

The response of users to these applied fees and credits is yet to be determined—however the application of these fees could be tested in a field trial. One significant benefit of the continuous collection of usage data is that we can alter the pricing pattern in a stepwise fashion, until the desired outcome is obtained. If users are responsive to fees, then the imbalances are solved without cost to user—as their behavior change in using the bikes would reduce the physical imbalances in the system. If users do not respond to the fee structure in terms of changing demand, then the cost of rebalancing would be applied to the users that as a collective create the structural imbalance. Users with cheap patterns of travel (using from and to stations without an imbalance) would be excused from any additional cost—and in our test case, that was women to a high degree.

Conclusions
Large scale data offers the policy analyst the opportunity to fine tune the practices and standards of modern transportation systems in close to real time. In addition, the continuous collection of data allows the policy makers to adjust in a more fine grained fashion and progressively address system challenges through a number of minor actions as opposed to gross actions.

Today, if we compare our road or transit system data as compared to bikeshare systems, it is obvious that the bikeshare systems are feeding back useful and instructive data on usage and fee payment. In stark contrast, our road and transit systems provide little trip level data that has any demographic or social data that would allow policy makers to tailor their solutions to promote environmental justice. Our results to date indicate that there are significant environmental justice issues with respect to gender and income in the bikeshare systems studied. As such, we encourage policy makers to examine this detailed data on a regular basis to address the needs of protected classes.
On exploring spatial correlations in census derived socioeconomic datasets

Vinay Uday Prabhu$^{1,2}$, Rohit Negi$^1$ and Miguel Rodrigues$^3$

$^1$Electrical and Computer Engineering Department, Carnegie Mellon University
$^2$Departamento de Ciencia de Computadores, University of Porto
$^3$Electronic and Electrical Engineering Department, University College London

Email: vinayprabhu@cmu.edu, negi@ece.cmu.edu, m.rodrigues@ucl.ac.uk
Abstract

In this paper, we use openly extractable census derived socioeconomic datasets to demonstrate two things. First, the successful non-linear regression of the Median Household Income (MHI) of 57 counties of the state of California using regressor variables such as Overall Poverty Level (OPL), Uninsured Population Level (UPL), Childhood Poverty Exposure Rate (CPER) and Adult Unemployment Level (AUL). Secondly, we also provide a parsimonious Gauss Markov Random Field (GMRF) model for the residuals in order to tackle the issue of underestimation of variance which would arise if we are to ignore the spatial correlation amongst the residuals across counties with a simplistic independence assumption.

Keywords: Spatial regression, Gauss-Markov Random Fields (GMRF), Variance Estimation.
1 Introduction

Supervised spatial regression has been hitherto performed in a wide array of contexts in the area of social sciences (Ward & Gleditsch, 2008). This includes domains such as disease mapping (Wakefield, 2007), Dendroclimatology (Cook, Briffa, & Jones, 1994), wide-area ecological studies (Mauricio Bini et al., 2009), Agriculture (Bongiovanni & Lowenberg-DeBoer, 2001), child poverty spread (Voss, Long, Hammer, & Friedman, 2006) and Forestry (Kaimowitz, Mendez, Puntodewo, & Vanclay, 2002).

Many of these endeavors are carried out in a temporal context usually involve training a regressor, linear (Neter, Wasserman, & Kutner, 1989) or non-linear (Bates & Watts, 1988), using legacy data and performing regression on a near-future dataset where the learned model parameters would still be relevant. When such an exercise is carried in a spatio-geographical context, where each sample pertains to a particular geographical jurisdiction, such as a county or a census tract, we encounter an intriguing problem. Training of the regressor is done with the assumption that the samples are statistically independent. This assumption ignores the rather rich spatial correlation that exists between the samples which brings us to the following question. In scenarios where the regressor performs admirably well in terms of standard regression error metrics such as Normalized Error (NE), in spite of ignoring the spatial correlation between the samples, can one still incorporate the spatial correlation and improve the regression analysis being performed? The answer, we believe, involves bringing in a statistical prior into the residuals (or noise), which will result in improved error variance estimation. This will make the model more amenable to handling systemic errors which can be completely missed with the independent and identically distributed (i.i.d.) noise assumption (Zhu, Ghosh, & Goodwin, 2009). Thus, one can say the regressor is entrusted with extracting all the intrinsic (intra-sample) information that exists in a sample while the extrinsic (inter-sample) spatial correlation information is handled by the spatial prior of the residuals and used to better estimate noise variance.

Underestimating noise variance in general on account of ignoring correlations carries dire consequences in many areas of policy design. Underestimated variances will lead to narrower confidence intervals for metrics such as the average probability of financial default (Coppens, González, & Winkler, 2007) or result in underestimated wage premia for fatality and injury risk in compensation analysis (Garen, 1988). In the area of Portfolio Decision Making, (Agarwal & Naik, 2004) have illustrated the serious extent to which the traditional mean-variance framework (Levy & Markowitz, 1979) underestimates the tail risk of hedge funds. In areas such as health policy, variance underestimations might lead to erroneous understanding of the impact of prenatal care on birth-weight (Todd Jewell & Triunfo, 2006).
In this paper, we specifically look at spatial county-wise regression of the Median Household Income (MHI) using regressor variables such as poverty levels, uninsured populace levels and unemployment rates. We use a non-linear Support Vector Regressor (SVR) with a Radial Basis Function (RBF) kernel function and showcase the under-estimation of variance that occurs by assuming that the regressions residuals are statistically independent across counties and also demonstrate that using a GMRF model for the residuals results in better variance estimation. The rest of the paper is organized as follows. In Section 2, we motivate the utility and importance of MHI as a policy metric and explain the dataset used for demonstrating the results. In Section 3, we present the regression model and introduce the GMRF model for the residuals. In Section 4, we present the methodology used and the results obtained. In Section 5, we comment on the novelty of this contribution and provide the future course of research. We also provide an appendix to cover the specific details of training the SVR.

2 Regression of Median Household Income: Motivation and dataset description

The socioeconomic metric of MHI has been heavily used as a critical criterion to shape and implement policy devising procedures for both urban and rural geographical regions. The U.S. Environmental Protection Agency (EPA) has used MHI in devising community-level drinking water pricing regulations (Rubin, 2001). The local city government of Houston uses the low-income household criterion (as defined by the U.S. Department of Housing and Urban Development (HUD) as being 80% or below the regional median household income) to decide the eligibility of an applicant to access special needs housing (Center, 2014). The U.S. Department of Transportation (DOT, 2003) use MHI to evaluate the value of personal travel for surface vehicles. MHI has also been used to help identify the largest vulnerable population that will be at risk for residing in a so-called food desert (Rubin, 2001). Therefore, driven by the magnitude of importance given to the metric in the above detailed literature, we chose MHI as the regression parameter of choice for our analysis.

2.1 Dataset description

In this paper, we are focusing on regression of the MHI of 57 counties\(^1\) of the state of California (CA) Figure 1, from 2007 through to 2012, mined from openly available public

\(^1\)CA has 58 counties. The county of Yuba was removed from our analysis on account of paucity of data with regard to the regressor variables.
databases, ("USDA, Economic Research Service", 2014) and ("Factfinder2 - Census.gov", 2014). The regressor variables used are: Overall Poverty Level (OPL), Uninsured Population Level (UPL), Childhood Poverty Exposure Rate (CPER) and Adult Unemployment Level (AUL), all expressed as probabilities and each mined from the openly available census data ("Factfinder2 - Census.gov", 2014). The framework presented here can be extended to other similar scenarios involving Urban data, where the counties can be replaced by Urban census tracts and the regression variables can be any other relevant socioeconomic metrics such as those related with public health, crime, race dynamics etc.

3 Model for Regression and residuals

Let \( n \) denote the number of counties and let \( f \) denote the number of regressor feature variables. For a given \( i^{th} \) county, the regression model we have is,

\[
y_i = g(x_i^T) + w_i, \tag{1}\]

where \( y_i \in \mathbb{R} \) is the dependent regressed variable, \( x_i \in \mathbb{R}^f \) is the vector of \( f \) regressor variables (features), \( g() \) is the regression function and \( w_i \in \mathbb{R} \) is the error (residual) variable, assumed to be Gaussian. Written in vector form, we have,

\[
y = \psi(X) + w, \tag{2}\]

where \( y \in \mathbb{R}^n, \psi(X) = [g(x_1^T); ..., g(x_n^T)] \in \mathbb{R}^{n \times f}, w \in \mathbb{R}^n \). Also, \( w \sim \mathcal{N}(0, \Sigma_w) \), where \( \Sigma_w \in \mathbb{R}^{n \times n}_+ \) is the (positive-definite) noise covariance matrix. The specific form of regression function \( g() \) chosen in this paper is the kernelized \( \varepsilon \)-SVR (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997) with RBF Kernel of a certain optimized bandwidth.

3.1 The GMRF model for noise

The main focus of the paper lies in enhanced modeling the covariance of the residuals (noise), \( w \). In most machine learning applications, the residuals are assumed to be i.i.d, that is,

\[
w \sim \mathcal{N}(0, \sigma^2 I_n), \tag{3}\]

where \( \sigma^2 \) is the noise variance.

We instead argue that in the case of the spatio-economic dataset under consideration, this i.i.d. assumption is unrealistic. We propose using the underlying spatial contiguity based
undirected (geographic) graph of counties (Figure 1), to construct a Gauss Markov Random Field (GMRF) model (Rue & Held, 2005) for the noise. That is,

\[ w \sim N(\mathbf{0}, \Sigma), \]  

where \( \Sigma \) is the GMRF covariance matrix. The inverse covariance matrix, \( \Omega = \Sigma^{-1} \) is parameterized as a homoskedastic conditional autoregression (CAR) model (Rue & Held, 2005), with the following parsimonious parameterization:

\[ \Omega = k (\mathbf{I}_n - \theta \mathbf{A}), \]  

where \( k > 0 \) is the conditional precision parameter, \( \theta \) is the homogeneous edge-weight parameter\(^2\), \( \mathbf{A} \) is the adjacency matrix of the undirected spatial graph, \( G(V, E) \), with \( V \) being the vertex set of counties (with \( |V| = n \)) and \( E \) being the edge-set. This model implies that the distribution of noise at node \( i \), with neighboring vertex set \( N_i \), conditioned on the rest

\(^2\)assumed to be positive given the homophilic nature of socio-economic tendencies across neighboring counties
of the variables is given by,

\[ w_i | w_{-i} \sim N \left( \theta \sum_{j \in N_i} w_j, k^{-1} \right), \]  \hspace{1cm} (6)

where \( w_{-i} \) is the vector \( w \) with the \( i^{th} \) element deleted. As seen, the inverse covariance matrix will be sparse with non-zero entries only on the diagonals and off-diagonal entries pertaining to the edge-locations. That is,

\[ \Omega_{i,j} = \begin{cases} 
-k \theta & \text{if } i \neq j, (i,j) \in E \\
 k & \text{if } i = j \\
 0 & \text{otherwise} 
\end{cases} \]  \hspace{1cm} (7)

4 Methodology and results

4.1 Training the SVR

To begin with, we use the \( t^{th} \) year’s data, \( D^{(t)} = \left\{ \left( x_1^{(t)}, y_1^{(t)} \right), \ldots, \left( x_n^{(t)}, y_n^{(t)} \right) \right\} \) and train the SVR. The SVR needs specification of three input parameters, namely, the RBF Kernel bandwidth parameter \( \gamma \), the loss function insensitivity width parameter, \( \varepsilon \), and the SVR-cost parameter, \( C \). The significance of these parameters is explained in detail in the Appendix. \( \gamma \) and \( \varepsilon \) were chosen via a 2-fold hold out cross validation procedure. The SVR-cost parameter, \( C \), was set to be the range of the regressed variable,

\[ C = y_+^{(t)} - y_-^{(t)}, \]  \hspace{1cm} (8)

where \( y_+^{(t)} = \max_i \left\{ y_i^{(t)} \right\} \) and \( y_-^{(t)} = \min_i \left\{ y_i^{(t)} \right\} \).

Once the SVR is learned, the residue (error) vector, \( \hat{w}^{(t)} \), is simply, \( \hat{w}^{(t)} = y^{(t)} - \hat{y}^{(t)} \). Figure 2 shows the histogram of regression errors (residuals) for the year 2009.

4.2 Parameter estimation for the Gaussian noise models

For the i.i.d. scenario, the unbiased estimated noise variance is simply,

\[ \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left( w_i^{(t)} - \hat{\mu} \right)^2, \]  \hspace{1cm} (9)
where \( \hat{\mu} \) is the estimated mean, \( \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} (w^{(t)}_i) \).

For the GMRF case, we have to estimate the parameter vector, \( \nu = [k \ \theta] \). The maximum likelihood estimates are derived by solving the following convex optimization problem, (Rue & Held, 2005),

\[
\nu^* = \max_{\nu: \Omega(\nu) > 0, \nu \geq 0} \left\{ \log \det (\Omega(\nu)) - tr (S^{(t)} \Omega(\nu)) \right\},
\]

(10)

where \( \Omega(\nu) \) denotes the positive definite inverse covariance matrix, \( \Omega \), populated as per (7) for every \( \nu \) chosen and \( S^{(t)} = \hat{\mathbf{w}}^{(t)} (\hat{\mathbf{w}}^{(t)})^T \). Also, \( \det() \) refers to the determinant of the matrix and \( tr() \) refers to the trace operation.

The GMRF covariance matrix is then, simply,

\[
\Sigma^{(t)} = \Omega^{-1}(\nu^*).
\]

(11)

Table 1 shows the values of the ML estimates for \( k^* \) and \( \theta^* \), obtained by training on years 2007 through to 2011.
Training Year | $k^*$ | $\theta^*$
---|---|---
2007 | 1.7550 | 0.1616
2008 | 1.6710 | 0.1414
2009 | 1.4511 | 0.1616
2010 | 1.4560 | 0.1818
2011 | 1.5278 | 0.1818

Table 1: ML estimates, $k^*$ and $\theta^*$, for different years

4.3 Testing on subsequent year’s data and variance estimation

The learned SVR is used to regress the following year’s regression variable using that year’s feature matrix, $X^{(t+1)}$. That is,

$$\hat{y}_i^{(t+1)} = g\left(x_i^{(t+1)}\right); i = 1, ..., n.$$ (12)

Figure 3 shows the true and regressed values of the MHI vector $y$ with the training year being 2009 and the testing year being 2010. We can ascertain the quality of the regressed variables $\hat{y}$ by evaluating the normalized error (NE), which is,

$$NE^{(t)} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{|y_i^{(t)} - \hat{y}_i^{(t)}|}{y_i^{(t)}} \right\}$$ (13)

Figure 4 shows the NE obtained by SVR with RBF kernel in regressing the MHI obtained for different years. As seen, the normalized error rates obtained are acceptably low and lie in the 9% to 11% range. Also shown in Figure 5 are the associated $r^2$-Squared correlation coefficient values defined as:

$$r^2 = \frac{\left[ n \sum_{i=1}^{n} \left\{ y_i^{(t)} \hat{y}_i^{(t)} \right\} - \sum_{i=1}^{n} y_i^{(t)} \sum_{i=1}^{n} \hat{y}_i^{(t)} \right]^2}{\left[ n \sum_{i=1}^{n} (y_i^{(t)})^2 - (\sum_{i=1}^{n} y_i^{(t)})^2 \right] \left[ n \sum_{i=1}^{n} (\hat{y}_i^{(t)})^2 - (\sum_{i=1}^{n} \hat{y}_i^{(t)})^2 \right]}.$$ (14)

As seen, the $r^2$ values obtained are in the acceptable range of 0.77 to 0.85 further vindicating the SVR’s regression performance.

Now, in order to showcase the fact that the GMRF approach indeed estimates variance better than the i.i.d. approach, ideally, we will need a lot of years’ worth of data. Instead, we use the following procedure which obfuscates the need for more data by creating sub-datasets within the given year’s dataset. To begin with, we divide the state into $n_{reg}$ regions. These can either be from naturally pre-existing regional classifications (Panel, 2009), as shown in...
Figure 3: True and regressed values of the MHI vector $y$. (Training year: 2009 — Testing year: 2010)

Figure 6, or we can define a region to be a node and its immediate neighbors. This way, using these 2 approaches, we get $n_{reg} = 9$ and $n_{reg} = 57$ sub-datasets respectively, to estimate the variance over and average to provide a conclusive comparison. Let $V_r$ denote the set of nodes in region $r$. The residual for that region, $\Delta_r$, is now defined to be,

$$
\Delta_r = \left( \sum_{j \in V_r} y_j^{(t+1)} \right) - \left( \sum_{j \in V_r} \hat{y}_j^{(t+1)} \right).
$$

(15)

We now define the Normalized Average Variance (NAV) estimate for the i.i.d. case to be,

$$
\varsigma_{i.d} = \frac{\sum_{r=1}^{n_{reg}} \left( \frac{\Delta_r^2}{\hat{\sigma}^2} \right)}{n_{reg}},
$$

(16)

where $\hat{\sigma}^2 = (|V_r|) \hat{\sigma}^2$ is the estimated variance from the i.i.d assumption. Similarly, for the GMRF case, we have the NAV estimate to be,
Figure 4: Normalized error of the regressed MHI obtained for different years.

$$\zeta_{gmrf} = \frac{\sum_{r=1}^{n_{reg}} \left( \Delta r^2 \right)}{n_{reg}}$$

(17)

where $\tilde{\sigma}_r^2 = e^T \Sigma \big|_{V_r,V_r} e$ is the estimated variance emanating from the GMRF assumption and $\Sigma \big|_{V_r,V_r}$ is the sub-matrix of $\Sigma$ obtained by considering only those columns and rows pertaining to the nodes of the vertex subset $V_r$. (Here $e$ is the vector of all ones). Figure 7 and Figure 8 show the estimated NAV obtained for different years for the $n_{reg} = 9$ economic region based approach and the $n_{reg} = 57$ node-neighborhood based region definition approach respectively. Ideally, we would like the NAV obtained to be as close to 1 as possible. As seen in Figure 7 and Figure 8, the i.i.d. assumption in resulting in severally underestimated variances which results in $\zeta$ being far greater to 1 when compared to the NAV obtained by the GMRF approach $\tilde{\zeta}$. These figures showcase the fact that modeling the residuals using the GMRF approach results in more realistic estimates of the variances associated with the regressed variables.
Figure 5: $r^2$ (Squared correlation coefficient) of the regressed MHI obtained for different years.

5 Novelty of contribution and future research

In literature, we find examples (such as (Zhu et al., 2009)), where spatial Auto-regressive (AR) models are used to model residuals in conjunction with linear regression in order to reduce the sampling variance of the estimated conditional distribution of the county (or region of counties) whose regression parameter is to be estimated. Our contribution differs from previous literature in two respects. Firstly, we go beyond linear regression and show the applicability of GMRF-Modeling of the residuals when used with a non-linear regression technique. Secondly, we showcase the utility of the GMRF noise prior even when spatial effects enter the model in a more subtle way, unlike in (Zhu et al., 2009), where the spatial correlation is induced by the sweeping weather across counties. In our scenario, the spatial effect is subtly induced on account of factors such as human contact, migration, knowledge sourcing, talent availability etc. This revelation paves the path for extending this model to other such inter-geographical unit datasets such as census-tract data for Urban zones.
Appendices

A Training the $\varepsilon$-SVR

We begin by assuming that we have been given $n_{\text{train}}$ samples of training data,

$$D = \{((x_i^T, y_i), \ldots, (x_{n_{\text{train}}}^T, y_{n_{\text{train}}})\},$$

with the feature vectors being $x_i \in \mathbb{R}^f$ and the regression variable being $y_i \in \mathbb{R}$. The $\varepsilon$-SVR, (Drucker et al., 1997), produces a regression function $g(x^T)$ that minimizes Vapnik’s $\varepsilon$-insensitive loss function (Vapnik, 1995), given by,

$$L_\varepsilon(y) = \begin{cases} 
0 & \text{for } |g(x^T) - y| < \varepsilon \\
|g(x^T) - y| - \varepsilon & \text{otherwise,}
\end{cases}$$

Figure 6: The 9 economic regions of CA as specified by the California Regional Economies Project.
Figure 7: Normalized Average Variance (NAV) estimate obtained for different years (for 9 economic regions based approach).

where the user-chosen $\varepsilon$, controls the insensitivity width of Vapnik’s loss function. The standard (kernelized) primal formulation of the SVR is given by,

$$
\min_{\{v,b,\xi,\xi^*\}} \left[ \frac{1}{2} v^T v + C \sum_{s=1}^{n_{train}} \xi_s + C \sum_{s=1}^{n_{train}} \xi_s^* \right]
$$

Subject to:

$$
\begin{align*}
    v^T \phi(x_s) + b - y_i & \leq \varepsilon + \xi_s, \\
y_i - v^T \phi(x_s) - b & \leq \varepsilon + \xi_s^*, \\
\xi_s^*, \xi_s & \geq 0, s = 1, ..., n_{train}.
\end{align*}
$$

Here, $C > 0$, is the so-called SVR-cost parameter intended to fine-tune the trade off between training error versus model complexity. Plugging in a small $C$ will enhance the number of training errors while a very large $C$ will lead hard-margin SVM like behavior (Joachims, 2002). As a rule of thumb, it is advised ((Mattera & Haykin, 1999),(Cherkassky & Ma, 2004)) to set the value of $C$ to be equal to the range of the regression variable $y$. Therefore, we set,

$$
C = y_+ - y_-,
$$
Figure 8: Normalized Average Variance (NAV) estimate obtained for different years (for node-neighborhood based region definition approach).

where $y_+ = \max \{y_s\}_{s=1}^{n_{\text{train}}}$ and $y_- = \min \{y_s\}_{s=1}^{n_{\text{train}}}$.

$\varepsilon$, which controls the width of the $\varepsilon$-insensitive loss function, intuitively allows us to manipulate the number of support vectors selected to build the regression function. Larger $\varepsilon$ values will result in fewer support vectors being selected.

$\xi_+^s$ and $\xi_-^s$ are the slack variables representing the lower and upper constraints on the outputs of the system. $\phi(.)$ are the kernel feature mappings that need to be explicitly evaluated on account of the Kernel trick.

The dual problem is written as,

$$
\min_{\{\alpha, \alpha^*\}} \left[ \frac{1}{2} \sum_{s=1, t=1}^{n_{\text{train}}} [(\alpha_s^* - \alpha_s) \mathbf{K}(x_s, x_t)(\alpha_t^* - \alpha_t)] + \sum_{s=1}^{n_{\text{train}}} (-\alpha_s^* + \alpha_s) y_s + \varepsilon \sum_{s=1}^{n_{\text{train}}} (\alpha_s^* + \alpha_s) \right]
$$

Subject to : $\sum_{s=1}^{n_{\text{train}}} (-\alpha_s^* + \alpha_s) = 0$,

$0 \leq \alpha_s^*, \alpha_s \leq C, s = 1, \ldots, n_{\text{train}}.$

(22)
Here, $\alpha$ and $\alpha^*$ are the Lagrange multipliers and $K(x_s, x_t)$ is the RBF kernel function, which is defined as,

$$K(x_s, x_t) = \exp \left( -\gamma \|x_s - x_t\|^2 \right).$$  \hspace{1cm} (23)

The Kernel bandwidth parameter $\gamma$ as well as $\varepsilon$ were chosen via a 2-fold cross validation procedure.

Upon receiving a new feature vector $x$, the regression output is given by,

$$\hat{y}_i = \sum_{s=1}^{n_{\text{train}}} \left[ (\alpha^*_s - \alpha_s) K(x_s, x_i) + b \right].$$  \hspace{1cm} (24)

References


15 

Proc. NSF Workshop on Big Data and Urban Informatics, Nov. 2015 Page 978
Dynamic Agent Based Simulation of an Urban Disaster Using Synthetic Big Data

A.Yair Grinberger¹ Michal Lichter² and Daniel Felsenstein³

¹ asherya.grinberger@mail.huji.ac.il Department of Geography, Hebrew University of Jerusalem Mount Scopus, 91900 ISRAEL
² mlichter@gmail.com Department of Geography, Hebrew University of Jerusalem Mount Scopus, 91900 ISRAEL
³ msdfels@mscc.huji.ac.il Department of Geography, Hebrew University of Jerusalem Mount Scopus, 91900 ISRAEL

Abstract

This paper illustrates how synthetic big data can be generated from standard administrative small data. Small areal statistical units are decomposed into households and individuals using a GIS buildings data layer. Households and individuals are then profiled with socio-economic attributes and combined with an agent based simulation model in order to create dynamics. The resultant data is ‘big’ in terms of volume, variety and versatility. It allows for different layers of spatial information to be populated and embellished with synthetic attributes. The data decomposition process involves moving from a database describing only hundreds or thousands of spatial units to one containing records of millions of buildings and individuals over time. The method is illustrated in the context of a hypothetical earthquake in downtown Jerusalem. Agents interact with each other and their built environment. Buildings are characterized in terms of land-use, floor-space and value. Agents are characterized in terms of income and socio-demographic attributes and are allocated to buildings. Simple behavioral rules and a dynamic house pricing system inform residential location preferences and land use change, yielding a detailed account of urban spatial and temporal dynamics. These techniques allow for the bottom-up formulation of the behavior of an entire urban system. Outputs relate to land use change, change in capital stock and socio-economic vulnerability.

Keywords: Agent based simulation, earthquake, synthetic big data, socio-economic profiling

Acknowledgements: This research is partially based on work done in the DESURBS (Designing Safer Urban Spaces) research project funded by the European Commission FP7 Program under Grant Agreement # 261652.
1. Introduction

The routine management of cities requires information regarding population characteristics, infrastructure, land-use, house prices, commercial activity, and so on. In the advent of a hazardous event (both natural and man-made) these different, yet interrelated sub-systems demand an immediate response. Invariably, the data requirements for this are only available at a coarse spatial resolution, such as TAZs or statistical units. Different data may be available at varying scales and administrative divisions. Moreover, the spatial extent of the disaster will probably not neatly overlap these divisions. Therefore, urban disaster management requires data which is detailed and dynamic, spatially and temporally. While high resolution, big data for individuals is becoming increasingly available (such as geo-tagged social media data, mobile phones, GPS tracking etc) and free of arbitrary spatial configurations, these data often over represent eager-sharers and under represent the technologically-challenged. Furthermore these data often do not include crucial socio-economic profiling such as that provided by surveys.

In this paper we show how existing ‘small’ administrative data can be utilized to generate ‘synthetic’ urban big data. We emphasize that such synthetic data is essential for disaster management. Big data is generated by decomposing small areal units into households and individuals, profiling them with socio-economic attributes and combining this data with an agent based simulation model in order to create dynamics. We use a GIS buildings layer to disaggregate administrative small data into households, individuals and eventually to the level of the synthetic location of individuals within buildings by floors. The resultant data can be considered ‘big’ in terms of volume, variety and versatility. Potentially, it allows for different layers of spatial information to be populated and embellished with synthetic attributes. This process of decomposition facilitates moving from a database describing hundreds or thousands of spatial units to one containing records of millions of buildings and individuals over time. The result is a comprehensive set of spatial 'big data' in which every single individual in a city is synthetically represented by a set of socio-economic attributes and high frequency dynamics.

A popular approach to handling this synthetic data is to intersect it with hazard maps to create a visual dynamic account of the development of a disaster. We have done this elsewhere using a dynamic web-interface that combines flood-hazard maps with the socio economic attributes of the areas under threat (Lichter and Felsenstein 2012). Alternatively, this data can serve as input for dynamic agent based (AB) modeling of urban disasters. This paper illustrates the latter route. We use data fusion techniques at the level of the individual building to generate the initial starting population data for an agent based simulation of an urban disaster. These techniques allow for the bottom-up formulation of the behavior of an entire urban system.

The behavioral response of each agent is determined according to its socio-demographic profiling. For example age, income and car ownership may constrain, enable, or affect travel mobility, activity selection, and residential location choice. In this manner, the data animates the population analytics for a dynamic agent based

Proc. NSF Workshop on Big Data and Urban Informatics, Nov. 2015  Page 980
simulation of an earthquake in Jerusalem. The AB simulation is based upon individual
citizen agents and their interaction with the built environment and with each other.
Feeding off the disaggregated data, individual buildings are characterized in terms of
land-use, floor-space and value. Agents are characterized in terms of income and
socio-demographic characters and are allocated to residential buildings. Using simple
behavioral rules grounded in risk-evasiveness, satisficing behavior, and residential
location preferences, along with a dynamic house pricing system that informs land-
use dynamics, a detailed description of urban spatial and temporal dynamics is
presented.

The paper proceeds as follows. After reviewing AB modeling applications for
urban disasters, we outline the modeling framework and context of the study. We then
describe how the big data is generated and coupled with the AB model. Simulation
results are presented relating to change in land use, capital stock and socio-economic
structure of the study area. To embellish the visualization potential of this data, we
present some output in the form of dynamic web-based maps. Finally, we speculate
on further developments derived from this approach.

2. Literature Review; AB Modeling for Disaster Management

Agent based modeling provides an appropriate framework for disaster
management (Fiedrich and Burgdardt 2007). In an agent based world, autonomous
entities (agents) behave according to a set of pre-programmed and simplistic, decision
rules. The activities of multiple agents create a computable system in which the
actions of individual agents affect each other and the system as a whole. The result is
a complex network of behavior patterns that could not have been predicted by simply
aggregating individual agent behavior. More importantly, such a system can be
simulated and subjected to various exogenous shocks. It is no wonder therefore that a
whole slew of disaster management situations have been subjected to AB simulations
especially where human organization and learning patterns can be programmed into
the response behavior of agents. These applications range from flooding (Dawson,
Peppe and Wang 2011), wildfires (Chen and Zhan 2008), epidemics (Simoes 2012) to
hurricanes (Chen, Meaker and Zhan 2006) and earthquakes (Crooks and Wise 2013).
Much of this effort produces output that highlights the collective behavior of the
agents. This can range from simulating the intervention of first responders, predicting
human behavior under conditions of stress, anticipating traffic and infrastructure
congestion due to human movement patterns and even simulating assistance efforts,
post disaster.

Less attention has been paid to simulating the response of this behavior on the
built environment. At the micro level of individual buildings and human response
patterns, Torrens (2014) has shown how the use of highly granular models can yield
rich detail of building collapse, agent-agent interaction and evacuation dynamics in
the case of a simulated urban earthquake. This contrasts with the ‘dumb, coarse and
cursory’ (Torrens 2014, p.965) nature of other AB models that try to reconcile human
and physical processes. The spatial and temporal dynamics of such a situation that are
animated by an AB model give rise to a huge volumes of information that while not intuitively recognized as ‘big data’ certainly qualify as such in terms of volume and variety.\(^1\) At the broader, system-wide level of the urban area, Zou et al (2013) argue that the bottom-up dynamics of AB simulation become homogenized when looking at complex urban processes such as sprawl or densification. They propose a different simulation strategy to that commonly used in AB modeling. This involves ‘short-burst experiments’ within a meta-simulation framework. It makes for more efficient and accelerated AB simulation and allows for the easy transfer across different spatial and temporal scales. Elsewhere, we have also illustrated that complex macroscopic urban change such as land use rejuvenation and morphology change in the aftermath of an earthquake, can be suitably analyzed in an AB framework (Grinberger and Felsenstein 2014, 2015).

While the literature addressing urban outcomes of disasters and using agent-based modeling is limited, there is a larger ancillary literature that indirectly touches on this issue from a variety of traditions. Chen and Zhan (2008) use the commercial Paramics simulation system to evaluate different evacuation techniques under different road network and population density regimes. Another approach is to couple GIS capabilities to an existing analytic tool such as remote sensing and to identify disaster hot spots in this way (Rashed et al., 2007). Alternatively, GIS can be integrated with network analysis and 3D visualization tools to provide a real-time micro-scale simulation tools for emergency response at the the individual building or city block level (Kwan and Lee 2005). At a more rudimentary level, Chang (2003) has suggested the use of accessibility indicators as a tool for assessing the post disaster effects of earthquakes on transportation systems.

A particular challenge to all forms of simulation modeling comes from the inherent dynamics of AB simulation and the visualization of results that has to capture both spatial and temporal dimensions. In the context of big data, this challenge is amplified as real time processing needs to also deal with large quantities of constantly changing data. These are state of the art challenges that require the judicious use of computational techniques, relational databases and effective visualization. The literature in this area is particularly thin. In a non-AB environment, Keon et al (2014) provide a rare example of how such integration could be achieved. They illustrate an automated geocomputation system in which a tsunami inundation and the resultant human movement in its aftermath are simulated. They couple the simulation model with a web-based mapping capability thereby allowing the user specify input parameters of their choosing, run the simulation and visualize the results using dynamic mapping via a standard web browser. A mix of server-side and client-side programming is invoked that allows the user all the standard functionality of web-based mapping.

Our approach takes this integration one stage further. In addition to using AB simulation we do not just combine a simulation model with a web-based visualization capacity but also generate the synthetic big data that drives the model. Once derived, this data need to be spatially allocated. The literature provides various methods such

\(^1\)While not calling this ‘big data’ as such, Torrens (2014) notes that the volume of locations/vectors to resolve for each object moved in the simulation is of the order of $10^{15}$ - $10^{17}$.\(^2\)
as population gridding (Linard et al 2011), areal interpolation which calls for 'creating' locations (Reibel and Bufalino 2005) and dasymetric representation which uses ancillary datasets such as roads or night-time lights imagery to approximate population location (Eicher and Brewer 2001, Mennis 2003) An alternative approach, adopted here, is to distribute the data discretely without resorting to interpolation or dasymetric mapping by combining different sources of survey and administrative data to create population count data. The result is then spatially referenced creating local population count data by utilizing an appropriate spatial anchor. In spirit, this method is closest to Harper and Mayhew (2012a,b) where local administrative data are combined and geo-referenced by a local source of land and property data.

3. The Modeling Framework

As behoves big data, the modeling framework used here is data-driven. The process is outlined in Fig 1. Socio-economic data for coarse administrative units (small data) is disaggregated into buildings and individuals on the basis of a GIS building layer and then recombined into households. The resultant synthetic data gives an accurate socio-economic profiling of the population in the study area. Coupling this data with an AB simulation models adds both temporal and spatial dynamics to the data. The result is a multi-dimensional big data set that affords flexibility in transcending conventional administrative boundaries. Outputs relate to socio-economic change, change in land use and capital stock in the aftermath of an earthquake. To fully capture the richness of the data generated we use web-based mapping to generate extra visual outputs.

Figure 1: The modeling framework
The Context

We simulate an earthquake in downtown Jerusalem (Figure 2). While Jerusalem is located 30 km southeast of the active Dead Sea Fault line the last major earthquake in the city occurred in 1927. The city center lies in a relatively stable seismic area but many of its buildings were constructed prior to the institution of seismic-mitigation building codes making it prone to damage (Salamon, Katz & Crouvi, 2010). The study area houses 22,243 inhabitants, covers 1.45 sq km and is characterized by low-rise buildings punctuated by high rise structures. A heterogeneous mix of land uses exist represented by residential buildings (243Th sqm, 717 structures), commercial buildings (505Th sqm, 119 structures) and government/public use buildings (420Th sqm, 179 structures). The area encompasses two major commercial spaces; the Machaneh Yehuda enclosed street market and the CBD. Three major transportation arteries roads traverse the area and generate heavy traffic volumes: Agripas and Jaffa Streets (light railway route) run north-west to the south-east and King George Street runs north-south. The area exhibits a heterogeneous mix of residential, commercial, governmental and public land use and high traffic volumes.

Figure 2: Study area

1 – CBD, 2 – Machaneh Yehuda Market, 3 – Jaffa St., 4 – Agripas St., 5 – King George St.
4. Big Data for Big Disasters

4.1 Generating Synthetic Big Data

Accurate spatial representation of data is essential for dealing with emergency situations in real time and for preemptive emergency management and training. However, spatial phenomena often do not neatly overlap administrative units of analysis. A need exists for accurately distributing the alpha-numeric socio-economic data for individuals or households within the spatial units used for data collection. We use a GIS-based system that allocates populations into buildings. We then create a spatial database where each inhabitant is represented as a unique entity. Each entity is attached a suite of unique socio-economic properties. Together, all the personal entities in every original spatial unit accurately represent the distribution of all socio-economic variables in the original unit. The disaggregation and reallocation process allows for the accurate spatial distribution of social and economic variables not generally available. This synthetically generated big data drives the AB simulation and underpins the unique characteristics of the agents. These characteristics will determine their behavioral choices.

The model is driven by data at three different resolutions: buildings, households and individuals. The original data is provided in spatial aggregates called Statistical Areas (SA)\(^2\). This is the smallest spatial unit provided by the Israeli Central Bureau of Statistics (CBS) census\(^3\). We use a disaggregation procedure whereby spatial data collected at one set of areal units is allocated to a different set of units. This transfer can be effected in a variety of ways such as using spatial algorithms, GIS techniques, weighting systems etc (Reibel and Bufalino 2005). The GIS layer provides the distribution of all buildings nationally with their aerial footprint, height and land use. We derive the floor-space of each building and populate it with individuals to which we allocate the relevant socio-economic attributes of the SA to which they belong, according to the original distribution of these attributes in the SA. In this way, synthetic big data is created from spatial aggregates. The mechanics of the derivation are described in Appendix 1. The variables used to populate the buildings and drive the model are:

- Building level: land-use, floor-space, number of floors, building value, households
- Household level: inhabitants, earnings, car ownership
- Individuals level: Household membership, disability, participation in the work force, employment sector, age, workplace location.

\(^2\) A statistical area (SA) is a uniform administrative spatial unit defined by the Israeli Central Bureau of Statistics (CBS) corresponding to a census tract. It has a relatively homogenous population of roughly 3,000 persons. Municipalities of over 10,000 population are subdivided into SA’s.

\(^3\) We also use coarser, regional data on non-residential plant and equipment stock to calculate non-residential building value. The estimating procedure for this data is presented elsewhere (Beenstock, Felsenstein and Ben Zeev 2011).
The variables used in the model, their sources and level of disaggregation appear in Table 1.

Figure 3: Data processing stages
Table 1: Variables used in the model

<table>
<thead>
<tr>
<th>Variable</th>
<th>Source</th>
<th>Spatial unit</th>
<th>Value</th>
<th>Disaggregation level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residential building value per m²</td>
<td>National Tax authority 2008-2013</td>
<td>SA</td>
<td>2008-2013 averaged (2009 real) value in NIS per m²</td>
<td>Building</td>
</tr>
<tr>
<td>Non-residential plant value</td>
<td>Local authorities financial data</td>
<td>Local authority</td>
<td>Total value of non-residential building stock in NIS per region</td>
<td>Building</td>
</tr>
<tr>
<td>Non-residential machinery and equipment value</td>
<td>Estimation (Beenstock et al 2011)</td>
<td></td>
<td>Total value of non-residential equipment stock in NIS per region</td>
<td>Building</td>
</tr>
<tr>
<td>Number of households</td>
<td>CBS 2008</td>
<td>SA</td>
<td>Total number of households per SA</td>
<td>Building, household</td>
</tr>
<tr>
<td>Number of inhabitants</td>
<td>CBS 2008</td>
<td>SA</td>
<td>Total number of inhabitants per SA</td>
<td>Building, household, individuals</td>
</tr>
<tr>
<td>Average monthly household earnings</td>
<td>National Insurance Institute, annual data</td>
<td>Local authority</td>
<td>Average household monthly earnings by SA</td>
<td>Building, household</td>
</tr>
<tr>
<td>Labor force participation</td>
<td>CBS 2008</td>
<td>SA</td>
<td>Working / not working</td>
<td>Building, household, individuals</td>
</tr>
<tr>
<td>Employment by sector</td>
<td>CBS 2008</td>
<td>SA</td>
<td>Commercial / Governmental / Industrial / Home-based / Unknown</td>
<td>Building, household, individuals</td>
</tr>
<tr>
<td>% disabled</td>
<td>CBS</td>
<td>SA</td>
<td>% disabled in SA</td>
<td>Building, household, individuals</td>
</tr>
<tr>
<td>Age</td>
<td>CBS</td>
<td>SA</td>
<td>0-18 / 18-64 / 65+</td>
<td>Building, household, individuals</td>
</tr>
<tr>
<td>Workplace location</td>
<td>GPS survey 2014</td>
<td>survey of individuals</td>
<td>Inside the region / outside the region</td>
<td>individuals</td>
</tr>
</tbody>
</table>

Buildings Level disaggregation: The basis of the disaggregation procedure is calculating the floor-space of each building using height and land-use. We assume an average floor height of 5 m for residential building and 7 m for non-residential buildings (see Appendix 1). These figures are the product of comparing total national built floor-space (for each land-use) with total national floor-space as calculated from the building layer.

The entire data disaggregation procedure is automated using SQL and Python code and the results at each stage are stored in a spatial database. The process entails first allocating inhabitants into buildings and then assigning them socio-economic
attributes. Later, these inhabitants are grouped into households and a further allocation of households attributes is performed. This necessarily involves loss of data due to dividing whole numbers (integers) such as households and inhabitants by fractions such as building floor-space for density calculations and percentages for socio-economic attribute distributions. In order to avoid loss of data and to meet the national control totals in each calculation, the SQL code is written so that it compensates for data losses (or increases) in the transition from floating points to integer values and ensures that the original control totals are always met. This is done by adjusting the floating point figures rounding threshold for each variable separately, to fit the losses or gain in the count of variables automatically.

Disaggregation at the level of the individual: The disaggregated building level data serves as the basis for the further disaggregation at the level of the individual. The building database includes a total of 1,075,904 buildings. A total of 7,354,200 inhabitants are allocated to 771,226 residential buildings. Disaggregation of the data to the individual begins with assigning each individual in the database a unique id, so that it is represented as a unique separate entity tied to a building in the database. Next, each person is allocated a random point location (a lat, lon coordinate) within the building with which it is associated. In each building, demographic attributes (labor force participation, employment sector, disabilities and age group) are allocated to each individual so that they comprise the entire distribution in the building which in turn gets its distribution from the SA in which it is located. In the same way, the distribution of work locations of inhabitants by employment sector is derived from a GPS-based transport survey carried by the Jerusalem Transport Master Plan Team (Oliveira et al., 2011). This is used to determine the distribution of inhabitants working inside or outside the study area according to their sector of employment and to assign the corresponding binary values to individuals.

Household level clustering and attribute allocation: Individuals are clustered into households by size of the household in each building. This creates new unique entities in the database representing households. Households are associated with buildings, and inhabitants are assigned to them. The clustering introduces heterogeneity in terms of the age distribution in the household to closely represent a “traditional household” containing both adults and children when these are present. This is achieved by an algorithm iterating through inhabitants in each building sorted by age, and assigning them to households. Depending on the age distribution in the building, this algorithm clusters inhabitants in a building into closely age represented but not identical, households. Each household is assigned the SA average household earnings value. Other attributes such as car ownership are assigned to households in the same way.

4.2 Coupling the Data with the Agent Based Model

The high resolution data detailed above is combined with an agent-based model. As agents represent the focal catalysts of change and aggregate patterns are decomposed into actions of individual agents, this further unshackles the constraints imposed by data collected on the basis of arbitrary administrative borders. In the
context of the current study this allows us to relate to the specific spatio-temporal nature of the event and its long-term impacts. To do this we characterize the three basic elements of an AB model: agents, their environment and rules governing agent-to-agent and agent-to-environment interactions (Macal & North, 2005).

The data provides the first two with individuals and households as agents and buildings as the urban environment. The model itself reflects the dynamics of the urban system via a collection of simplistic behavioral rules. These govern the interactions within the system at each iteration (Fig 4) which is set to represent one day and also include an exogenous shock in the form of the earthquake. These rules are described in Appendix 2.

Figure 4: Conceptual formulation of agent-based model of an earthquake in a city

A simulation entity is created to represent each individual, household and building along with its spatial and socio-economic characteristics. Identifying the unique workplace for each employed individual in the study area is done on the basis of satisficing behavior. The first building which satisfies randomly generated preferences in terms of distance from home location and floor-space size (assuming larger functions attract more employees) and is of the land-use associated with the individual’s employment sector is designated as the agent’s workplace.
Agent behavior (bottom-up procedures): This simulation characterizes the city as a spatial entity whose organization emerges from the aggregate behavior of all its citizens (individuals). Individuals and households are therefore identified as the agents of the urban system. Their behavior is simplified into two decision sets: the decisions of households about place of residence and the decisions of individuals about choice and sequence of daily activities.

The decision to change place of residence is probabilistic and based on comparing a randomly drawn value to exogenous probabilities of migrating out of the study area (city-level probability), or relocating to another part of the study area (SA specific probability). Choosing a new place of residence location follows two decision rules: a willingness to commit up to one third of monthly household earnings to housing and preferences regarding the socio-economic characteristics of the residential environment. This follows a probabilistic and satisficing procedure similar to that described for selection of work place. If a household fails to find an alternative place of residence after considering 100 possible locations, it leaves the study area. Individual agents that are members of the household relocate/migrate with the household. In-migration is treated in the model by having the number of potential migrating households dependent on the number of available housing units and an exogenous in-migration/out-migration ratio. New households with up to 2 members are comprised of adults only. Those with more members include at least one non-adult agent and their socioeconomic status reflects the urban average for key socioeconomic attributes.

At each iteration individuals conduct a variety of activities within the study area. These are important for land-use dynamics and the mobility paths between land uses (see below). The number of activities undertaken ranges from 0-11 and varies by socio-economic characteristics (age, car ownership of household, disability, employment status, location of employment) and randomly generated preferences. The location of each activity (with the exception of work activity for agents employed within the study area) is determined by the attractiveness of different locations. This in turn is dependent on floor-space size, environment, distance from previous activity, and the mobility capability of the individual. This choice criteria is again, probabilistic and satisficing. Movement between each pair of activities is not necessarily shortest-path. A more simplistic aerial-distance-based algorithm is used to reduce computing demands and again reflect the satisficing nature of agents.

Environmental influences (top-down procedures): AB models typically have a demand side emphasis characterizing the urban system as the outcome of residents’ behavior. In normally functioning markets this means that agents will look for cheaper housing. On the supply side however, contractors will tend to build where prices (profits) are high and thus house prices will vary directly with population and

---

4 Calculated from 2012 immigration data in the Statistical Yearbook for Jerusalem 2014 (Jerusalem Institute for Israel Studies).

5 The algorithm works as follows: at each step, junctions adjacent to the current junction are scanned and the junction with the shortest aerial distance to the destination is flagged as the current junction for the next step. If a loop is encountered, it is deleted from the path. The algorithm ends when agents arrive at the junction closest to the destination or when all junctions accessible from the origin are scanned.
inversely with housing stock. In AB models the supply side is often overlooked. We formulate a top-down process in order to get a fuller picture of the operation of urban markets. We conceptualize administrative units, buildings and individual housing units as quasi-agents. These entities are not autonomous or mobile but are nevertheless sensitive to changes in their environment according to pre-defined rules of influence (Torrens 2014). Foremost amongst these are sensitivity to the distribution of commercial activity and to values of buildings. In this process, changes on the demand side (population dynamics), and related changes to the land-use system create effects which trickle down from the level of the administrative unit to the individual dwelling unit and ultimately to the individual resident.

We assume that the larger a commercial function, the more activity it will require in order to be profitable. We further assume that the volume of activity is represented by the intensity of traffic in the vicinity of each function. Consequently, traffic volume on roads near commercial functions should be proportional to their floor-space. Hence, any function that is not located near sufficient traffic will not survive and residential uses near large traffic volumes may change their function to commercial, forcing residents to relocate or migrate.

Changes to a dwelling unit’s value represent a top-down spatial mechanism based on supply and demand dynamics. Average price per meter is pre-specified at the beginning of the simulation for each SA, for both residential and non-residential uses. As land-use and population dynamics change, supply and demand shift causing these average prices to fluctuate. We assume that non-residential value is affected only by changes in the supply (short supply drives prices up) while residential values are also sensitive to changes in demand (prices drop when population decreases). Changing average prices cause a change in building values according to floor-space. In the case of residential buildings this change is also effected through accessibility to services. Accessibility relates to the share of non-residential uses from total uses accessible from the building. A building whose accessibility is higher than the SA average will be higher priced. These effects further trickle down to the monthly cost of housing per dwelling unit. We assume uniformity of unit prices within a building meaning that the cost of each unit is the average cost of units in the building. This cost affects household residential choice in the next iteration.

The earthquake: this is formalized as a one-time shock, spreading outwards from a focal point with intensity decaying over distance. This earthquake is programmed to occur on the 51th simulation iteration, so the system has a sufficient 'run-in' period in which all processes initiate themselves and stabilize. The epicenter of the quake is determined randomly so that aggregate average results do not represent any place-based bias. As the impact spreads across space it may inflict damage to the environment. The probability a building being demolished as a result of the shock is proportional to both distance from the epicenter and height. A demolished building causes the nearest road to become unavailable for travel until the building is restored to use. The restoration period is proportional to the building’s floor-space volume.
5. Simulation and Results

Results are presented relating to three main themes: land use change, change in value of capital stock and socio-economic change. These reflect three dimensions of urban vulnerability to unanticipated shocks: functional, economic and social vulnerability respectively. The simulation is run 25 times each with a duration of 1000 iterations (i.e. days). The earthquake occurs on day 51 in each simulation and is located randomly within the study area. The 50 day run-in time is heuristically derived, comprising of a 30 day period for stochastic oscillation and a further 20 day period for ‘settling down’. Like any catastrophic change, the initial impact of the earthquake is an immediate population loss and damage to buildings and infrastructure. Yet as the results illustrate, these lead to a second, indirect round of impacts induced by the way in which agents react to the new conditions created.

Traffic Patterns and Land Use Change: we present a series of maps that illustrate change in land use (from residential to commercial and vice versa) and the concomitant dispersal of traffic activity at discrete time points (Figures 5 & 6). As can be seen, in the period following the disaster the main commercial thoroughfares running N-S and E-W across the area lose their prominence induced by changing movement patterns. Within 50 days of the event, major traffic loads shift from the north-west part of the study area into the south and to a lesser extent to the north-east and central sections. Commercial activity responds to this change and a cluster of medium-sized commercial uses appears in the south-west. However, by day 500 we observe a reversal of this pattern. Evidently, the recovery of the traffic network, along with the anchoring effect of large commercial land uses, helps Agripas St. to regain its position causing a new commercial cluster to develop in its vicinity. The immediate, post disaster change in traffic pattern does however leave a permanent footprint. Commercial land use becomes more prevalent in the north-east and CBD centrality seems to be slightly reduced. The buildings containing these activities have larger floor-space area than those located in the new emerging commercial clusters. This implies a potentially large addition of dwelling units to the residential stock in the case that these buildings transform into residences. One year after the shock these patterns hardly change as the new transportation pattern gets locked in, empty buildings become occupied and the potential for land use change is reduced. From the 3D representation of frequency of land use change (Fig 6) we can identify the time juncture where buildings that were previously empty become occupied. Between day 500 and day 1000 land use tends to become rejuvenated in the sense that unoccupied buildings become populated.
Figure 5: Average land-use maps at discrete time points

Figure 6: Frequency of land-use change at discrete time points

Frequency is represented by building height. Building color represents initial use. Colored section represents the share of total simulations in which the building was in use other than the original use, at that time point. Grey represents the share of times the building was unoccupied. Road height represents absolute traffic volume, while shading represents relative volume within the distribution of all roads.
**Change in the Value of Capital Stock:** standard urban economic theory suggests that demand for capital stock (residential and non-residential) is inversely related to price while the supply of capital stock is positively related to price. In our AB world with dynamic pricing for both residential and non-residential stock, demand changes through population change and supply changes through either building destruction or change in land use as a result of changing traffic loads and accessibility to services. Aggregate simulation results for residential capital stock show that the number of buildings drops to about 600 after about 100 days and in the long run never recovers (Figure 7). However average residential values tend to rise only after 500 days to about 90 percent of their pre-shock values. This lagged recovery may be attributed to supply shortage and increased access to services as suggested by the changing land use patterns noted above along with increasing demand from a growing population. Yet, the fact that a reduced residential stock recovers to almost the same value suggests that this is due to rising average floor-space volumes. This would point to buildings that were initially large commercial spaces becoming residential. Non residential stock behaves rather differently. Long-term increase in stock is accompanied by lower average values. In contrast to residential stock, the turning point in these trends is after roughly one year (Figure 8). Elsewhere we have identified this with the dispersal of commercial activities from large centers to smaller neighborhood units (Grinberger and Felsenstein 2014). The current results reflect a similar picture. The number of units grows but their average value decreases as the buildings in the south-west and north-east have smaller floor space.

![Figure 7: Change in residential capital stock](image-url)
Population Dynamics: The initial impact causes a population loss of about 4,000 residents (Figure 9). After about one year population size recovers and continues to grow to about 29,000 by the end of the simulation period. This increase is the result of in-migration of an heterogeneous population with stochastic earnings. The ability of this extra population to find residence in the area is due to the process of land use change described above. The new, large residential buildings (previously commercial spaces) contain many individual dwelling units. While the average price of buildings rises, the rising average floor-space of buildings pushes down the average cost per dwelling unit within the building and consequently the monthly cost of housing services. As a result, lower income households find suitable lower cost residential location making for an increased population that is poorer on average. Since the average income of new in-migrants is slightly higher than the average income of the area, this suggests that the lower average income must result from the out-migration of wealthier households that accompanies in-migration.
A composite indicator of both social, functional and economic vulnerability can be obtained by looking at the flow-through of households through buildings. The simple ratio of in-coming to out-going households per building at each discrete time step, gives an indication of the amount of through-traffic per building and an indication of its population stability. Fig 10 gives a summary account of this ratio. A high ‘pull’ factor is not necessarily a sign of stability or even attractiveness. It may be an indicator of transience and instability. The overall picture is one of unstable population dynamics. The simulations suggest that none of the buildings that were initially residential are consistently attractive to population. Most of them have difficulty maintaining their population size, post earthquake. For many buildings this is due to either physical damage or change in function for example, from residential to commercial use. It seems that only new potential residential spaces that start initially as commercial uses, consistently succeed in attracting population. The direct and indirect effects of the shock generate much household turnover (or ‘churning’) through buildings but without any indications of location preferences for specific buildings or parts of the study area. Total floor space that registers net positive household movement (strong/weak pull) amounts to only 75%of the floor space that registers net negative turnover (strong/weak push). This underscores the high state of flux in the study area.
Social Vulnerability: a more in-depth examination of population movement is presented in Figure 11 where snapshots of the spatial distribution of social vulnerability at different time steps are presented. We follow Lichter and Felsenstein (2012) and use a composite index of social vulnerability\(^6\). Green indicates less vulnerable population and red more vulnerable. The average index value for each building is calculated, disaggregated into households and used to generate continuous value surfaces using Inverse Distance Weighting (IDW). The parameters used for the interpolation are: pixels of 10X10 meters, 100 meters search radius and a 2\(^{nd}\) order power function.

Akin to movement patterns, residence patterns in day 100 present a process of dispersal as relatively less vulnerable households initially clustered in the west move eastward. This population is higher income and consequently has greater availability of resources for relocation. Since agents are characterized by limited tolerance to change it is not surprising to find that the destinations chosen by households are those that were better-off in the pre-event situation (light green areas). Residential patterns re-cluster after day 100, a process similar to the stabilization of traffic and land-use patterns. However, this re-clustering germinates a new spatial pattern with small

\(^6\) Social vulnerability by household \((V_{hh})\) is defined as:

\[ V_{hh} = 0.5 \times Z_{ihh} - 0.2 \times Z_{agehh} - 0.2 \times Z_{%dishh} + 0.1 \times Z_{car} \]

where: \(Z\) is the normalized value of a variable, \(i\) is household income for household \(hh\), \(age\) is the average age group of members of household \(hh\), \(%dis\) is the percent of disabled members of all members in household \(hh\), \(car\) is car ownership for household \(hh\).
cluster of stronger population serving as nuclei for future agglomeration. These clusters attract households of similar characteristics resulting in the emergence of new clusters and a general process of entrenchment. This process occurs at a high spatial resolution that is much more granular than the SA level. Visualizing at this scale prevents the homogenization of patterns. For example, we can detect dispersion over time of the high vulnerability population as the east-west vulnerability patterns in the pre-shock era becomes replaced by a much more heterogeneous picture.

Figure 11: Social vulnerability heat map
6. Web-Based Delivery of Outputs

The richness of outputs is hard to envision and to communicate. The volume and variety of the big data used as input into the model is manifested in the complexity of these outputs. Not only are these produced for variables at different spatial units but they are also generated temporally for each variable. These dynamics are typically non-linear. The nature of the information generated necessitates a visualization technique capable of accommodating large amounts of data and presenting them spatially and temporally. This in turn means sophisticated database construction that will allow not only the dynamic representation of data but also the querying of data in a user-friendly fashion. We view this as a crucial linkage between the simulation outputs generated in a research oriented environment and their use by potential consumers such as planners, decision makers and the informed public.

The results are thus presented via a web-based application. This allows for communicating outcomes in a non-threatening and intuitive way (see http://ccg.huji.ac.il/AgentBasedUrbanDisaster/index.html). Using a web browser, users can generate results for different points in time and for different areas without prior knowledge of handling spatial data or GIS. They can choose a variable of interest, visualize its change over time and space and generate the relevant location specific information they need. To this end, we create a dedicated database for the output results of time series from the model simulation. This needs to be carefully constructed and sometimes does not follow strict DB design but rather contains some flat tables of lateral data in order to be displayed in pop-ups graphs and charts. The visualization includes time lapse representation of human mobility (household level), changes in passengers along roads, changes in buildings’ land use and value, household socio-economic attributes change etc. in the study area (Figures 12-13).

We use Google Maps API as the mapping platform and add middleware functionalities. These are functions that are not provided by the mapping API but interact with it to provide ancillary capabilities (Batty et al. 2010) such as time laps animation, sliders, interactive graphs etc. These middleware functionalities are User Interface (UI) features that allow for different ways of data querying and interactive engagement, using a variety of JavaScript libraries and API’s. Utilizing this mashup of visualization tools is merely the final stage in the development of the web-map. It is preceded first by extensive data analysis and manipulation of vast census and model output spatial data and second by creating a dedicated database in order to allow easy, intuitive and sometimes lateral querying of the database.
Figure 12: Time lapse visualization of various household socio-economic attributes on a dynamic web-map (see http://ccg.huji.ac.il/AgentBasedUrbanDisaster/index.html)

Figure 13: Time lapse visualization of the change in the number of passengers along roads on a dynamic web-map (see http://ccg.huji.ac.il/AgentBasedUrbanDisaster/index.html)
7. Conclusions

This paper makes both a methodological and substantive contribution to the study of urban dynamics using recent advances in urban informatics and modeling. In terms of method we illustrate how an agent based model can be coupled with a data disaggregation process in order to produce synthetic big data with accurate socio-economic profiling. This fusion adds both temporal and spatial dynamics to the data. The simulation model uniquely treats the built environment as a quasi-agent in urban growth. Consequently, more attention is paid to the supply side dynamics of urban change than generally practiced and the result is a modeling system with dynamic pricing and an active supply side. We also illustrate how outputs can be suitably communicated to practitioners and the informed public. Dynamic web-based mapping is used to enhance civic engagement and public awareness as to the possible implications of a large scale exogenous shock.

On the substantive side the results of the simulation highlight some interesting urban processes at work. We speculate about three of them and their implications for the ability of cities to rejuvenate in the aftermath of an unanticipated event. The first relates to the permanent effects of temporary shocks. Our results have shown that temporary shocks to movement and traffic patterns can generate longer term lock-in effects. In our simulations these have a structural effect on reduction of CBD commercial activity. The issue arising here is the ability to identify when this fossilization takes place and when a temporary shock has passed the point of no return.

The second process relates to the large level of household turnover and ‘churning’ through the built fabric of the city in the aftermath of an earthquake. Obviously, a traumatic event serves to undermine population stability as housing stock is destroyed and citizens have to find alternative residence. However, high turnover levels of buildings point to a waste of resources, material, human and emotional. In other markets such as the labor market, ‘churning’ might be considered a positive feature pointing to economic and occupational mobility (Schettkat 1996). However, in the context of a disaster, this would seem to be a process that judicious public policy should attempt to minimize. The welfare costs of the effort needed to search for new accommodation and the dislocation associated with changing place of residence are likely fall hardest on weaker and more vulnerable populations (Felsenstein and Lichter 2014).

Finally, our findings shed new light on the familiar concept of ‘urban vulnerability’. Our simulated results show that less vulnerable socio-economic groups ‘weather the storm’ by dispersing and then re-clustering over time. This points to their higher adaptive capacities. Stronger populations have the resources to accommodate the negative impacts of a disaster. Urban vulnerability is thus as much as an economic welfare issue as it an engineering or morphological concept. From a socioeconomic perspective, it is not the magnitude of the event that is important but the ability to
cope with its results. This makes urban vulnerability a relative term: a shock of a
given magnitude will affect diverse population groups differentially. Vulnerable
populations or communities can be disproportionately affected by unanticipated
disasters which are more likely to push them into crisis relative to the general
population. Much of this can only be detected at the micro level such as the
household. It is often smoke-screened in studies dealing with aggregate citywide
impacts. The use of highly disaggregated and accurately profiled data would seem to
be critical in understanding urban vulnerability.
Appendix 1: Data Disaggregation Method

This appendix describes the disaggregation procedure of spatially aggregated alpha numeric real-estate values and populations and their socio-economic attributes into discrete spatial units at the building level.

The number of floors in residential buildings \( (F_R) \), is calculated by dividing building height by average floor height of 5 m:

\[
F_R = \frac{H_B}{5}
\]

In the case of non-residential buildings, the number of floors \( (F_N) \) is estimated as the building height divided by average floor height of 5 m:

\[
F_N = \frac{H_B}{7}
\]

Floor space for each building \( (S_B) \) is then calculated by multiplying the number of floors in each building by its polygon area representing roof space:

\[
S_B = S_R \times F
\]

Where:
- \( S_R \): Building polygon footprint
- \( F \): Building number of floors

The GIS buildings layer and building type serve as the basis for the calculation of residential building value, non-residential building and equipment value. To create estimates of residential building value we use average house prices per m\(^2\) 2008-2013 (in real 2009 prices). In cases where no transactions exist in a specific SA over that period we use a regional estimate for residential property prices.

1. **Value of residential buildings** \( (P_{BR}) \) is calculated as follows:

\[
P_{BR} = P_{SR} \times S_{BR}
\]

where:
- \( P_{SR} \): Average SA price per m\(^2\)
- \( S_{BR} \): Residential building floor space.

2. **Value of Non residential buildings** is calculated as follows:

Non residential value per m\(^2\) by region \( (P_{RN}) \):

\[
P_{RN} = \frac{V_{RN}}{S_{RN}}
\]

where:
- \( S_{RN} \): Total regional non residential floor space.
- \( V_{RN} \): Total regional non residential building stock

Non residential building value per m\(^2\) for each region is multiplied by the floor space of each non-residential building to produce non-residential building values \( (P_{BN}) \):

\[
P_{BN} = P_{RN} \times S_{BN}
\]

where:
- \( S_{BN} \): non residential building floor space.
Regional non-residential stock estimates have been calculated for nine aggregate regions elsewhere (Beenstock et al. 2011).

3. **Value of Equipment and Machinery** ($P_{RE}$) is calculated as follows:

$$P_{RE} = \frac{V_{RE}}{S_{RN}}$$

where:
- $V_{RE}$ = Total regional non-residential equipment stock

The equipment stock per m$^2$ for each region is multiplied by the floor space of each non-residential building to produce equipment stock totals by building ($P_{BE}$):

$$P_{BE} = P_{RE} \times S_{BN}$$

where:
- $S_{BN}$ = non-residential building floor space.

The source for regional estimates of regional equipment and machinery stock is as above (Beenstock et al. 2011).

The buildings layer also allows for the spatial allocation of aggregated households, and population counts (see Table 1) into building level households and inhabitant totals. Given the existence of these spatial estimates, the distribution of aggregate average monthly earnings, participation in the work force, employment sector, disabilities and age (see Table 1) into a building level distribution is implemented.

4. **Household density by SA (households per m$^2$)** of residential floor space in a statistical area ($H_{SR}$) is calculated as follows:

$$H_{SR} = \frac{H_{S}}{S_{SR}}$$

where:
- $H_{S}$ = Total population per statistical area (IV in Table 1).
- $S_{SR}$ = Total statistical area residential floor space.

The number of households per building ($H_B$) is calculated as follows:

$$H_B = H_{SR} \times S_{BR}$$

5. **Average number of inhabitants per m$^2$** of residential floor space in a statistical area ($I_{SR}$) is calculated as follows:

$$I_{SR} = \frac{I_{S}}{S_{SR}}$$

where:
- $S_{SR}$ = Total statistical area residential floor space.
- $I_{S}$ = Total population per statistical area.

Population counts per building ($I_B$) are then calculated as follows:

$$I_B = I_{SR} \times S_{BR}$$

6. **Total earnings per building** ($M_B$) is calculated as follows:
Where:

\[ M_B = M_{SI} \times H_B \]

where:
- \( M_{SI} \): Average monthly earnings per household by SA.
- \( H_B \): Total number of households in a building.

7. **Number of inhabitants in each building participating in the labor force 2008 (I_w)** is calculated by multiplying the number of inhabitants in a building by the labor participation rate in the corresponding SA.

\[ I_w = W_S \times I_B \]

where:
- \( W_S \): % of inhabitants participating in the labor force in an SA
- \( I_B \): Population count per building

8. **Number of inhabitants per building by employment sector (I_o)** is calculated by multiplying the percentage of inhabitants employed by sector (commercial, governmental, industrial or home-based) per statistical area by the number of inhabitants in each building.

\[ I_o = O_S \times I_B \]

where:
- \( O_S \): % of inhabitants employed in an employment category.
- \( I_B \): Population counts per building

9. **Number of disabled inhabitants in each building (I_d)** is calculated by multiplying the number of inhabitants in a building by the percentage of disabled in the corresponding SA.

\[ I_d = D_S \times I_B \]

where:
- \( D_S \): % of disabled inhabitants in an SA
- \( I_B \): Population count per building

10. **Number of inhabitants in each age category (I_a)** is calculated by multiplying the number of inhabitants in a building by the percentage of inhabitants in each age category in the corresponding SA.

\[ I_a = A_S \times I_B \]

where:
- \( A_S \): % of inhabitants in each age group category in an SA
- \( I_B \): Population count per building.
Appendix 2: Behavioral Rules for the ABM

1. Residential Location Choice is derived as follows:

\[ h_h = b_j \Rightarrow \left[ \frac{I_h}{3} > HP_j \right] \ast \left[ k_h > S(b_j) \right] = 1 \]

where:

- \( h_h \) is the new residential location for household \( h \),
- \( b_j \) is the building considered,
- \([ \ ]\) is a binary expression with value of 1 if true and 0 otherwise,
- \( I_h \) is household \( h \)'s monthly income,
- \( HP_j \) is monthly housing cost of an average apartment in building \( j \),
- \( k_h \) is a random number between \([0,1]\) indicating tolerance to change in residential environment incurred by relocation,
- \( S(b_j) \) is a similarity score for building \( j \) in relation to current place of residence, calculated as follows:

\[ S(b_j) = \frac{\Phi \left( \frac{I_j - I_h}{I_{\sigma_h}} \right) + \Phi \left( \frac{A_j - A_h}{A_{\sigma_h}} \right)}{2} \]

where:

- \( \Phi \) is the standard normal cumulative probability function,
- \( I_j, A_j \) are the average household income and average age of individuals in building \( j \), respectively
- \( I_h, A_h \) are average household income and average age of individuals in residential buildings within 100 meter of current residential location of household \( h \), respectively
- \( I_{\sigma_h}, A_{\sigma_h} \) are standard deviations of household income and of resident age in residential buildings within 100 meters from current home location of household \( h \), respectively.

2. Choice of sequence of activities: occurs in two stages. First, the number of activities is fixed and then activities are allocated to locations:

\[ \# Ac_i = \left| a \ast \left( \frac{k_i}{0.5} \right) \ast (1 + car_h \ast 0.33) \ast (1 - dis_i \ast 0.33) \ast (1 + [age_i = 2] \ast 0.33) \ast (1 - [age_i \neq 2] \ast 0.33) \right| + employed, \ast here_i \]

where:

- \( \# Ac_i \) is the number of activities for resident \( i \),
- \( k_i \) is a randomly drawn number between \([0,1]\) reflecting preferences regarding number of activities,
- \( car_h \) is a binary variable equal to 1 if the household \( h \) owns a car and 0 otherwise,
- \( dis_i \) is a binary variable equal to 1 if individual \( i \) is disabled and 0 otherwise,
- \( age_i \) is the age group of individual \( i \),
- \( employed \) is a binary variable equal to 1 if the individual \( i \) is employed and 0 otherwise,
- \( here \) is a binary variable equal to 1 if the location is within 100 meters from current home location and 0 otherwise.
employed, is a binary variable equal to 1 if \(i\) is employed and 0 otherwise, here, is a binary variable equal to 1 when \(i\)'s workplace is located within the study area and 0 otherwise, \(\|x\|\) indicates the nearest integer number to \(x\),
a is the average number of activities based on employment status; equals 2.5 for employed residents and 3 for non-employed.

\[
a_{i,t+1} = b_j \iff [b_j \neq a_{i,t}] * [k_i \geq \text{Attr}(b_j)] = 1
\]

where:
\(a_{i,t}\) is the current location of individual \(i\),
\(a_{i,t+1}\) is the next location of activity of individual \(i\),
\(k_i\) is a randomly drawn number between \([0,1]\) reflecting activity location preferences,
\(\text{Attr}(b_j)\) is the attractiveness score for building \(j\), calculated as follows:

\[
\text{Attr}(b_j) = \frac{1 - \sum_{i} E_j / \sum_{j} B_j + 1 - D_{ij} / \max D_i * (1 + 0.33 * (-\text{car}_{i,j} + \text{dis}_{i,j} + [\text{age}_i = 3])) + \text{LU}_j = \text{nonRes} * \frac{FS_j}{\max FS}}{2 + \text{LU}_j = \text{nonRes}}
\]

where:
\(\sum E_j\) is the number of non occupied buildings within a 100 meter buffer of building \(j\),
\(\sum B_j\) is the number of all buildings within a 100 meter buffer of building \(j\),
\(D_{ij}\) is the distance of building \(j\) from the current location of individual \(i\),
\(\max D_i\) is the distance of the building within the study area furthest away from the current location of individual \(i\),
\(\text{LU}_j\) is the land-use of building \(j\),
\(\text{nonRes}\) is non-residential use,
\(FS_j\) is the floor-space volume of building \(j\),
\(\max FS\) is the floor-space volume of the largest non-residential building within the study area.

3. Choice of workplace location is calculated similarly to the choice of activity location:

\[
WP_i = b_j \iff [\text{LU}_j = \text{ELU}_i] * [k_i > \frac{D_{ij}}{\max D_i} + 1 - \frac{FS_j}{\max FS}] = 1
\]

where:
\(WP_i\) is the workplace location of individual \(i\),
\(\text{ELU}_i\) is the employment-sector-related land-use for individual \(i\),
\(k_i\) is a randomly drawn number between \([0,1]\) representing workplace location preferences,
$D_{ij}$ is the distance between building $j$ and individual $i$’s place of residence, 

$max D_i$ is the distance of the building within the study area furthest away from individual $i$’s place of residence.

4. **Building values and the monthly cost of living in a dwelling unit** are derived in a 3-stage process. First, daily change in average house price per SA is calculated. Then, values of individual buildings are derived and finally the price of the single, average dwelling unit is calculated. For non-residential buildings, the calculation of individual building values is similar.

$$AHP_{z,t+1} = AHP_{z,t} \left( 1 + \log \left( \frac{pop_{z,t+1}}{pop_{z,t}} + \frac{res_{z,t}}{res_{z,t+1}} + \frac{nRe{s}_{z,t+1}}{nRe{s}_{z,t}} \right) \right)$$

$$ANRV_{z,t+1} = ANRV_{z,t} \left( 1 + \log \left( \frac{nRe{s}_{z,t}}{nRe{s}_{z,t+1}} \right) \right)$$

where:

- $AHP_{z,t}$ is average housing price per meter in SA $z$ at time $t$,
- $pop_{z,t}$ is population in SA $z$ at time $t$,
- $res_{z,t}$ is the number of residential buildings in SA $z$ at time $t$,
- $nRe{s}_{z,t}$ is the number of non-residential buildings in SA $z$ at time $t$,
- $ANRV_{z,t}$ is the average non-residential value per meter in SA $z$ at time $t$,

$$HP_{j,t} = AHP_{z,t} \cdot FS_j \cdot \frac{SL_{j,t}}{SL_{z,t}}$$

$$V_{j,t} = ANRV_{z,t} \cdot FS_j$$

where:

- $HP_{j,t}$ is the house price of a dwelling unit in building $j$ at time $t$,
- $SL_{s,t}$ is the service level within area $s$ at time $t$ – the ratio of non-residential buildings to residential buildings in this perimeter,
- $V_{j,t}$ is the non-residential value of building $j$. 

$$\bar{I}_t = \left( \frac{HP_{j,t}}{\sum_{l=1}^{n} HP_{i,l}} - \frac{\sum_{l=1}^{n} HP_{l,t}}{\sum_{l=1}^{n} \Sigma Ap_l} \right)$$

$$P_{du,t} = \left( \frac{1 + \frac{\sum_{l=1}^{n} \Sigma Ap_l}{P_{\sigma}}}{c} \right)$$
where:

\( P_{du,t} \) is the monthly cost of living in dwelling unit \( du \) at time \( t \),

\( \bar{I}_t \) is the average household income in the study area at time \( t \),

\( \Sigma Ap \) is the number of dwelling units within a building. If the building is initially of residential use, this is equal to its initial population size, otherwise it is the floor-space volume of the building divided by 90 (assumed to be average dwelling unit size in meters),

\( L_t \) is the number of residential buildings in the study area at time \( t \),

\( P_{\sigma} \) is the standard deviation of dwelling unit prices within the study area at time \( t \),

\( c \) is a constant.

5. Land-use changes, from residential to commercial and from commercial to unoccupied are based on the congruence between the building floor-space volume and the average intensity of traffic on roads within a 100 m radius over the preceding 30 days. Both these values are compared with the (assumed) exponential distribution of all values in the study area. This is done by computing the logistic probability of the relative difference in their locations in the distribution:

\[
P_{j,t}(\Delta x_{j,t}) = \frac{e^{-\Delta x_{j,t}}}{1 + e^{-\Delta x_{j,t}}}
\]

\[
\Delta x_{j,t} = \frac{z_{IR_{j,t}} - z_{FS_{j,t}}}{|z_{FS}|}
\]

\[
z_{y_{j,t}} = \frac{e^{y_{med_t}/\bar{y}_t} - e^{y_{j,t}/\bar{y}_t}}{\bar{y}_t}
\]

where:

\( P_{j,t} \) is the probability of land-use change for building \( j \) at time \( t \),

\( \Delta x_{j,t} \) is the relative difference in position of traffic load and floor-space for building \( j \) at time \( t \),

\( z_{y_{j,t}} \) is the position of value \( y \) in the exponential distribution, relative to the median for building \( j \) at time \( t \),

\( \frac{e^{y_{med_t}/\bar{y}_t}}{\bar{y}_t} \) is the exponential probability density value for \( y \) (\( \frac{1}{\bar{y}_t} = \hat{\lambda}_t \)) for building \( j \) at time \( t \),

\( y_{med_t} \) is the median of \( y \) at time \( t \).

If \( P>0.99 \) for residential use, it changes to commercial. If the value is in the range \([P(1)-0.01,P(1)]\) for commercial uses, the building becomes unoccupied. This functional form and criteria values reduce the sensitivity of large commercial uses and small residential uses to traffic volume. Consequently, the process of traffic-related land-use change is not biased by a tendency to inflate initial land uses.
6. *Earthquake impact* is calculated as follows:

\[ \text{Im}_j = \frac{c \times 10^{\text{mag}_j}}{D_j \times \log(D_j) \times F_j} \]

where:

- $\text{Im}_j$ is the impact building $j$ suffers, $c$ is a constant, $\text{mag}_j$ is the earthquake magnitude (similar to Richter scale), $D_j$ is distance of building $j$ from the earthquake epicenter, $F_j$ is number of floors in building $j$. 
References


Using Social Media to Task Data Collection and Augment Observations in Urban Areas During Emergencies: 2013 Boulder Floods Case Study

Guido Cervone∗,a,b, Emily Schnebelea, Nigel Watersa, Jeff Harrisond, Martina Moccaldi,a,c, Rosa Sicignanoa,c

∗GeoInformatics and Earth Observation Lab, Dept. of Geography and Inst. for CyberScience
The Pennsylvania State University, University Park PA

bResearch Application Laboratory, National Center for Atmospheric Research, Boulder CO

cDept. of Computer Science, University of Salerno, Italy
dThe Carbon Project, Arlington VA

Abstract

Environmental hazards pose a significant threat to urban areas due to their potential catastrophic consequences affecting people, property and the environment. Remote sensing has become the de-facto standard for observing the Earth and its environment through the use of air-, space-, and ground-based sensors. Numerous sensors are routinely used to collect massive amounts of dynamic and geographically distributed spatio-temporal data which are often used for damage assessment and mitigation. However, despite the quantity of data available, gaps are often present due to the specific limitations of the instruments, their carrier platforms, or as a result of atmospheric interference. Today, massive amounts of data are generated from social media, and it is possible to mine these data to fill the gaps in remote sensing observations. A new methodology is described which uses social networks for tasking the collection of remote sensing imagery of transportation infrastructure conditions in response to emergencies. The capability is valuable in situations where environmental hazards such as hurricanes or severe weather affect very large areas. During these types of disasters it is paramount to ‘cue’ remote sensing imagery collection to assess the impact of fast-moving and potentially life threatening events on transportation infrastructure.

This research presents an application of the proposed methodology during the 2013 Colorado floods with a special emphasis in Boulder County and The City of Boulder. Real-time data collected from social media, such as Twitter, identify and ‘cue’ where damage is being discussed. Commercial satellite and Unmanned Aerial Vehicles are tasked for the collection of high resolution imagery for damage assessment of the areas identified as being potentially compromised. The collected imagery is made available to a broad community of users and advisors through online services. Furthermore, data collected from social media can also provide information when remote sensing data are lacking or unavailable.

Key words: Remote Sensing, Social Media, Natural Disasters

1. Introduction

Every year natural hazards are responsible for powerful and extensive damage to people, property, and the environment. Drastic population growth, especially along coastal areas or in developing countries, has increased the risk posed by natural hazards to large, vulnerable populations at unprecedented levels (Tate and Frazier, 2013). Furthermore, unusually strong and frequent weather events are occurring worldwide, causing floods, landslides, and droughts affecting thousands of people (Smith and Katz, 2013). A single catastrophic event can claim thousands of lives, cause billions of dollars of damage, trigger a global economic depression, destroy natural landmarks, render a large territory uninhabitable, and destabilize the military and political balance in a region (Keilis-Borok, 2002). Furthermore, the increasing urbanization of human society, including the emergence of megacities, has led to highly interdependent and vulnerable social infrastructure that may lack the resilience of a more agrarian, traditional society. In urban areas, it is crucial to develop new ways of assessing damage in real-time to aid in mitigating the risks posed by hazards. Annually, the identification, assessment, and repair of damage caused by hazards requires thousands of work hours and billions of dollars. In 2011, over 300 natural disasters were reported, with victims totaling over 200 million and economic damage estimated at...
over $360 billion.¹

Remote sensing data are of paramount importance during disasters and have become the de-facto standard for providing high resolution imagery for damage assessment and the coordination of disaster relief operations (Cutter, 2003; Joyce et al., 2009). Organizations such as the International Charter for Space and Disasters² provide high resolution imagery from commercial and research air- and space-borne instruments within hours of major events, frequently including ‘before’ and ‘after’ scenes of the affected areas (Stryker and Jones, 2009; Duda and Jones, 2011). These ‘before’ and ‘after’ images are quickly disseminated through news channels to inform the public of the magnitude of the event, and often serve to sensibilize citizens about the unfolding tragedies. In addition, first responders rely heavily on remotely sensed imagery for coordination of relief and response efforts as well as the prioritizing of resource allocation.

Despite the wide availability of large remote sensing datasets from numerous sensors, specific data might not be collected in the time and space most urgently required. Geo-temporal gaps result due to satellite revisit time limitations, atmospheric opacity, or other obstructions. Tasking instruments on-board satellites and other aerial platforms for data collection is thus crucial for the timely delivery of data for damage assessment and disaster relief. However, satellite tasking is usually limited by orbital restrictions and the locations of data receiving stations. It is usually predefined, and based on the statistical likelihood that data for an area are needed. A small number of satellite instruments can be oriented to collect data at an oblique angle with respect to the satellite path. For this class of instruments, a correct tasking can greatly increase the data coverage during emergencies.

Furthermore, aerial platforms, especially Unmanned Aerial Vehicles (UAVs), can be quickly deployed to collect data about specific regions. UAVs are capable of providing high resolution, near real-time imagery often with less expense than manned aerial- or space-borne platforms. Their quick response times, high maneuverability and resolution make them important tools for disaster assessment (Tatham, 2009). Tasking data collection for specific regions most affected by a hazard is particularly difficult during major events, possibly compromising the larger region and testing the data collection capacity of limited systems. For extremely large events such as Hurricane Sandy in 2012 or the Colorado floods of 2013, rapid and systematic evaluations were difficult because the area affected was so extensive. This difficulty is further enhanced when events quickly unfold and are mitigated by local infrastructure and relief efforts.

Therefore, it is apparent that during emergencies the tasking of data collection from remote sensing platforms must be constantly assessed and refined based on the needs of emergency responders and the constraints dictated by the number and type of instruments available. This assessment has historically been based on official measurements and established plans, and did not account for the availability of real-time, on-the-ground data freely contributed by citizens.

Novel information streams, such as social media contributed videos, photos, and text as well as other open sources, are redefining situation awareness during emergencies. When these contributed data contain spatial and temporal information they can provide valuable Volunteered Geographical Information (VGI), harnessing the power of ‘citizens act as sensors’ to provide a multitude of on-the-ground data, often in real time (Goodchild, 2007). Although these volunteered data are often published without scientific intent, and usually carry little scientific merit, it is still possible to mine mission critical information. For example, during hurricane Katrina, geolocated pictures and videos searchable through Google provided early emergency response with ground-view information. These data have been used during major events, with the capture, in near real-time, of the evolution and impact of major hazards (De Longueville et al., 2009; Pultar et al., 2009; Heverin and Zach, 2010; Vieweg et al., 2010; Acar and Muraki, 2011; Verma et al., 2011; Earle et al., 2012; Tyshchuk et al., 2012).

Volunteered data can be employed to provide timely damage assessment, help in rescue and relief operations, as well as the optimization of engineering reconnaissance (Laituri and Kodrich, 2008; Dashi et al., 2014; Schnebele and Cervone, 2013; Schnebele et al., 2014a, b). While the quantity and real-time availability of VGI make it a valuable resource for disaster management applications, data volume, as well as its unstructured, heterogeneous nature,

¹http://www.emdcat.be/database
²http://www.disastercharter.org/
make the effective use of VGI challenging. Volunteered data can be diverse, complex, and overwhelming in volume, velocity, and in the variety of viewpoints they offer. Negotiating these overwhelming streams is beyond the capacity of human analysts. Current research offers some novel capabilities to utilize these streams in new, groundbreaking ways, leveraging, fusing and filtering this new generation of air-, space- and ground-based sensor-generated data (Oxendine et al., 2014).

This research presents a novel approach to prioritizing the collection and analysis of remote sensing data during hazard events by utilizing VGI as a filtering tool as well as information source to fill in the gaps when remote sensing data are lacking or incomplete. In order to efficiently and effectively use social media to ‘cue’ or augment satellite observations, it is necessary to filter the data for content and geolocate them using a variety of text-mining and network analysis algorithms. Filtering yields a rapid and directed identification of affected areas which can aid authorities in prioritizing site visits and response initiatives as well as the tasking of additional data collection. The present research proposes a methodology to task satellite and UAV data during emergencies based on social media and other open sources as well as the application of social media to augment a hazard assessments. In particular, we present an application of this new methodology to the 2013 floods that occurred around the city of Boulder, CO.

2. Data

Multiple sources of contributed, remote sensing, and open source geospatial data were collected and utilized during this research. A summary of the sources and collection dates of the contributed and remote sensing data is available in Table 1.

2.1. Contributed data

2.1.1. Twitter

The social networking site Twitter is used by the public to share information about their daily lives through microblogging. These micro-blogs, or ‘tweets’, are limited to 140 characters, so abbreviations and colloquial phrasing are common, making the automation of filtering by content challenging. Different criteria are often applied for filtered and directed searches of Twitter content. For example, a hashtag is an identifier unique to Twitter and is frequently used as a search tool. The creation and use of a hashtag can be established by any user and may develop a greater public following if it is viewed as useful, popular, or providing current information. Other search techniques may use keywords or location for filtering, like Gazetteers. In addition to tweets harvested specifically for this research (Section 3.2), Twitter data were also provided by The Pennsylvania State University (PSU) GeoVISTA center. The tweets were filtered by three different criteria:

- tweets extending from 105.0814° – 105.2972° W longitude and 40.09947° – 39.95343° N latitude (location: near Boulder) and containing the hashtag “#boulderflood” from 12 - 16 September 2013;
- tweets extending from 105.0322° – 105.4302° W longitude and 40.09947° – 39.93357° N latitude (location: near Boulder) and containing all hashtags from 11 -17 September;
- tweets extending from 12 - 16 September 2013 (all locations) and containing the hashtag “#boulderflood”.

2.1.2. Photos

Photos (n=80) which documented flooding in the Boulder area, from 11 - 14 September 2013 and extending from 105.07° – 105.32° W longitude and 40.12° – 39.97° N latitude, were collected using the Google search engine.

2.2. Remote sensing

2.2.1. Satellite

Ten, full-resolution GeoTIFF Worldview-2 multispectral images collected by Digital Globe on 13 September (2 images), 14 September (5 images), and 17 September (3 images) 2013 provide high resolution data of Boulder and the surrounding counties. Worldview-2 collects 8 bands of imagery at a 1.84m resolution: coastal (0.40-0.45µm), blue (0.45-0.51µm), green (0.51-0.58µm), yellow (0.585-0.625µm), red (0.63-0.69µm), red edge (0.705-0.745µm), near-IR 1 (0.77-0.895µm), and near-IR 2 (0.86-0.90µm). Using the application developed for this research, Worldview-2 data were collected for 13 September.
2013 and made available at our data portal. The remaining Worldview data were freely available by Digital Globe and were downloaded from the USGS Hazards Data Distribution System (HDDS). Full-resolution GeoTIFF multispectral Landsat 8 OLI-TIRS images that were collected on May 12, 2013 and on September 17, 2013 provide data of the Boulder County area before and after the flooding, respectively. The data were downloaded from the USGS Hazards Data Distribution System (HDDS). Landsat 8 consists of 9 spectral bands with a resolution of 30m: Band 1 (coastal aerosol, useful for coastal and aerosol studies, 0.43-0.45µm); Bands 2-4 (optical, 0.45-0.51, 0.53-0.59, 0.64-0.67µm), Band 5 (near-IR, 0.85-0.88µm), Bands 6 and 7 (shortwave-IR, 1.57-1.65, 2.11-2.29µm) and Band 9 (cirrus, useful for cirrus cloud detection, 1.36-1.38µm). In addition, a 15m panchromatic band (Band 8, 0.50-0.68µm) and two 100m thermal IR (Bands 10 and 11, 10.60-11.19, 11.50-12.51µm) were also collected from Landsat 8 OLI-TIRS.

2.2. Aerial

Aerial photos collected by the Civil Air Patrol (CAP), the civilian branch of the US Air Force, captured from 14 - 17 September 2013 in the areas surrounding Boulder (105.536° – 104.9925° W longitude and 40.26031° – 39.93602° N latitude) provide a third source of data. The georeferenced Civil Air Patrol RGB composite photos were downloaded from the USGS Hazards Data Distribution System (HDDS).

Unmanned aerial vehicles (UAVs) flown by Falcon Unmanned were deployed over the Lyons and Longmont, CO areas from 12 - 14 September, 2013. Falcon UAV collected valuable RGB composite photos while other aircraft were grounded due to weather conditions. The mosaiced, georeferenced images were downloaded as a .kmz file from the Falcon Unmanned website and are available at the project’s data portal.

2.3. Open source geospatial data

Shapefiles defining the extent of the City of Boulder and Boulder County were downloaded from the City of Boulder® and the Colorado Department of Local Affairs® websites, respectively. In addition, a 2012 TIGER/line® shapefile of road networks for Boulder County was downloaded from the US Census Bureau. All data layers were georeferenced to UTM coordinates in ArcGIS 10.1.

3. Methods

3.1. Overview

The proposed methodology is based on the application of contributed data (Twitter) to identify the location of natural hazards, or ‘hot spots’, and the tasking of remote sensing data for these areas. Additional ground data were then collected and integrated with the available remote sensing data for analysis and damage assessment (Figure 1). This damage assessment can be further utilized to classify transportation infrastructure during and after an event.

3.2. Remote sensing data collection

3.2.1. Identify hot spots

For this research, a scanning application was developed to access and harvest social media data by browsing Twitter in real-time. Tweets are filtered by keyword and hashtag, locating road, bridge and natural hazard condition reports. As events progress, the service ‘scans’ the United States each hour to assess and generate alerts for areas with significant Twitter activity. These alerts or ‘hot spots’, are areas where natural hazards may be potentially occurring, and are identified based on the clustering of filtered and, subsequently, geolocated tweets. When a ‘hot spot’ is identified, remote sensing data are

---

4 http://portal.cubewerx.com/DOT/
5 http://hddsexplorer.usgs.gov
6 http://www.falconunmanned.com/
7 http://portal.cubewerx.com/DOT/
8 https://bouldercolorado.gov
9 http://www.colorado.gov
10 http://www.census.gov
9.1.1

Table 1: Sources and collection dates of contributed and remote sensing data.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Contributed:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Twitter</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Photos</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Remote Sensing:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Worldview-2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Landsat-8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Civil Air Patrol</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Falcon UAV</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Sources and collection dates of contributed and remote sensing data.

collected and made available as open data services. The application system is highly flexible with filter settings managed by a portal service enabling manager-level users to quickly adjust keyword and other settings as events develop.

The geolocation of filtered tweets is accomplished by pairing geographic names identified in the tweet text to a United States gazetteer (Figure 2). Geographic names collected from the United States Geological Survey (USGS) were downloaded into a single database file - storing the name and the geometry properties of each named location. In cases where there are multiple points, the center point of the extent is utilized as the geometry. The resulting geonames database contains approximately 2.1 million records. The components are capable of running and reading Twitter streams and then matching the identified place names to the geonames database in real-time.

3.2.2. Task remote sensing

Following the filtering and geolocation of Tweets, a potential natural hazard event is recognized when a significant amount of activity is identified in a region. In this work, the threshold for significant activity is set to ten tweets within a 100km² area. A 100km² extent was selected because it is sufficient to capture significant transportation infrastructure in many urban areas as well as being the minimum ordering area for DigitalGlobe orthorectified imagery. The threshold settings can be adjusted to user requirements or preferences. When the threshold is reached, an ‘Alert Box’ is generated, cueing the collection of imagery for that region (Figure 3). The imagery is deployed in near real-time using OpenImageMap services as well as other open mapping services including Open Geospatial Consortium Web Map Service (OGC WMS), Web Map Tile Services (WMTS), Web Coverage Services (WCS), Google Maps API, Google Earth KML overlays, and Open Source Geospatial Foundation Tile Map Service (OSGeo TMS). Providing the imagery in an open source format ensures rapid deployment as well as open access to the data. System settings are adjustable and managed by a portal service enabling manager-level users to quickly add imagery, manage services, and access controls. Imagery may also be accessed by the WMS standard and combined with National Spatial Data Infrastructure (NSDI) framework data in any GIS supporting this popular standard.

3.3. Additional data collection and processing

Once a potential natural hazard has been identified by the scanning application and remote sensing data for

---

11http://www.cubewerx.com/solutions/openimagemap/
Figure 3: Example of Alert Boxes triggered when tweets reach a threshold of (n=10) within 100km².

the area have been tasked, additional sources of readily available ground data are collected to augment the analysis and assessment of the event. Data processing is source dependent and will vary depending on hazard type.

3.3.1. Classification of satellite images

For the Colorado floods of 2013, supervised machine learning classifications were employed to identify water in each of the satellite images.

Water pixels are identified in both Landsat images by using a decision tree induction classifier. Ripley (2008) describes the general rule induction methodology and its implementation in the R statistical package used in this study. In the near-IR water is easily distinguished from soil and vegetation due to its strong absorption (Smith, 1997). Therefore, imagery caught by Landsat’s Band 5 (near IR, 0.85-0.88µm) were used for the machine learning classification. Control areas of roughly the same size are identified as examples of water pixels, ‘water’, and over different regions with no water pixels as counter-examples, or ‘other’. Landsat data relative to these regions are used as training events by the decision tree classifier. The learned tree is then used to classify the remaining water pixels in the scene. This process is repeated for both the May and September images.

Maximum likelihood supervised classifications of the Worldview-2 images were conducted using ERDAS Imagine software. The supervised classification was applied to each scene individually to account for differences in brightness values and then 3 daily scenes (13, 14, and 17 September 2013) were created by mosaicing the classified images together for each date in ArcGIS 10.1.

3.3.2. Processing of aerial images, photos and tweets

Satellite remote sensing data may be insufficient as a function of revisit time or obstructed due to clouds or vegetation. Therefore, data from other sources can be used to provide supplemental information. Aerial remote sensing data as well as contributed data, or VGI, from photos and tweets were used to capture or infer the presence of flooding in a particular area. The methodology for deploying these data consists of two steps:

1) geo-location of data;
2) spatial interpolation of data.

Geo-location

In this paper, the geo-location of Twitter data are accomplished using two different methods. In the first example, tweets collected by the scanning application developed for this research are geolocated by matching place names garnered from the tweet text to a gazetteer database (Section 3.2.1). The second set of Twitter data, collected by The Pennsylvania State University’s GeoVISTA center, analyzes tweets for identifiers such as place names, hashtags, and organizations and then georeferences these locations using GeoNames12 (MacEachren et al., 2011).

The contributed photos were manually geolocated in ArcGIS 10.1 using location information included in the photos’ metadata such as street addresses or place names (i.e. Boulder Public Library).

The aerial RGB composite photos collected by Falcon Unmanned and the Civil Air Patrol came with geolocation information.

Spatial interpolation

Utilizing different types and sources of data, this research aims to extract as much information as possible about damage caused by natural hazards. Environmental data are often based on samples in limited areas, and the tweets analyzed are approximately only 1% of the total tweets generated during the time period. This is usually referred to the ‘Big Data paradox’, where very large amounts of data to be analyzed are only a small sample of the total

\[^{12}\text{http://www.geonames.org/}\]
data, which might not reflect the distribution of the entire population.

In addition, the absence of data in some parts of the region is likely to underestimate the total damage. In order to compensate for the missing data, we analyze the spatio-temporal distribution of the data, weighting accordingly the spatial relationship of points (Tobler, 1970, pg.236). This allows us to assume some levels of dependence among spatial data as well as to closely examine spatial information found to be inconsistent with its surroundings (Sui, 2004). For these reasons a punctual representation of data may not be sufficient to provide a complete portrayal of the hazard, therefore a spatial interpolation is employed to estimate flood conditions and damage from point sources.

Spatial interpolation consists of estimating the damage at unsampled locations by using information about the nearest available measured points. For this purpose, interpolation generates a surface crossing sampled points. This process can be implemented by using two different approaches: deterministic models and statistical techniques. Even if both use a mathematical function to predict unknown values, the first method does not provide an indication of the extent of possible errors, whereas the second method supplies probabilistic estimates. Deterministic models include IDW (Inverse Distance Weighted), Rectangular, NN (Natural Neighbourhood) and Spline. Statistical methods include Kriging (Ordinary, Simple and Universal) and Kernel. In this project, Kernel interpolation has been used.

**Kernel interpolation**

Kernel interpolation is the most popular non-parametric density estimator, that is a function \( \hat{p} : \mathbb{R} \times \mathbb{R}^N \rightarrow \mathbb{R} \). In particular it has the following aspect:

\[
\hat{p}(x) = \frac{1}{Nh} \sum_{i=1}^{N} K \left( \frac{x - x_i}{h} \right).
\]

where \( K(u) \) is the Kernel function and \( h \) is the bandwidth (Raykar and Duraiswami, 2006). There are different kinds of kernel density estimators such as Epanechnikov, Triangular, Gaussian, Rectangular. The density estimator chosen for this work is a Gaussian kernel with zero mean and unit variance having the following form:

\[
\hat{p}(x) = \frac{1}{N \sqrt{2\pi h^2}} \sum_{i=1}^{N} e^{-\frac{(x-x_i)^2}{2h^2}}. \tag{2}
\]

Kernel interpolation is often preferred because it provides an estimate of error as opposed to methods based on radial basis functions. In addition, it is more effective than a Kriging interpolation in case of small data sets (for example, the data set of photos in this project) or data with non-stationary behavior (all data sets used in this work) (Mühlenstädt and Kuhnt, 2011). In general, spatial interpolation is introduced to solve the following problems associated with histograms:

- The wider the interval, the greater the information loss;
- Histograms provide estimates of local density (points are “local” to each other if they belong to the same bin) so this method does not give prominence to proximity of points;
- The resulting estimate is not smooth.

These problems can be avoided by using a smooth kernel function, rather than a histogram “block” centered over each point, and summing the functions for each location on the scale. However, it is important to note that the results of kernel density interpolations are very dependent on the size of the defined interval or “bandwidth”.

**Bandwidth of kernel interpolation**

The result of a kernel density estimation will depend on the kernel \( K(u) \) and the bandwidth \( h \) chosen. The former is linked to the shape, or function, of the curve centered over each point whereas the latter determines the width of the function. The choice of bandwidth will exert greater influence over an interpolation result than the kernel function. Indeed, as the value of \( h \) decreases, the local weight of single observations will increase.

Because confidence in data may vary with source characteristics, bandwidth selection can be varied for each data type. The basic idea is that the more certain the information given by a kind of data, the higher chosen bandwidth. For example, all aerial images concern with certainly and considerably damaged areas. By contrast some tweets could not be useful since they are subjective; indeed some of them could only contain users’ feelings.
and not information related to damages caused by the hazard.

Therefore, the first kind of data provides more significant information than the second one. For this reason a lower bandwidth for tweets has been chosen in this work.

There are different methods for choosing an appropriate bandwidth. For example, it can be identified as the value that minimizes the approximation of the error between the estimate \( \hat{p}(x) \) and the actual density \( p(x) \) as explained in (Raykar and Duraiswami, 2006).

**Spatial kernel density estimation**

In this work, a spatial Kernel estimation has been used because all considered data sets consist of points on the Earth’s surface. In case of \( d \)-dimensional points the form of the Kernel estimate is:

\[
\hat{p}(x; H) = n^{-1} \sum_{i=1}^{N} K_H(x - X_i)
\]

where \( x = (x_1, x_2, \ldots, x_d)^T \), \( X_i = (X_{i1}, X_{i2}, \ldots, X_{id})^T \), \( i = 1, 2, \ldots, N \) and \( K_H(x) = |H|^{-1/2} K(H^{-1/2} x) \). In this case \( K(x) \) is the spatial kernel and \( H \) is the bandwidth matrix, which is symmetric and positive-definite. As in the uni-dimensional case, an optimal bandwidth matrix has to be chosen, for example, using the method illustrated in (Duong and Hazelton, 2005). In this project, data have been interpolated by using the R command smoothScatter of the package graphics based on the Fast Fourier transform. It is a variation of regular Kernel interpolation that reduces the computational complexity from \( O(N^2) \) to \( O(N) \). Generally, the bandwidth is automatically calculated by using the R command bkde2D of the R package KernSmooth. However, the bandwidth for tweets interpolation has been specified because information related to them has a lower weight.

4. **Analysis and Results**

4.1. **Damage Assessment**

Following the identification of a ‘hot spot’ of Twitter activity in Colorado beginning 11 September 2013, remote sensing imagery was tasked for the areas of interest. Using supervised machine learning classification as discussed in Section 3.3.1, water pixels were identified in these images. For example, a comparison of the classifications in the Landsat 5 May, 2013 ‘before’ image (Figure 4a) and the Landsat 17 September, 2013 ‘after’ image (Figure 4b) illustrates additional water pixels classified in the ‘after’ image associated with the flood event. One of the challenges associated with classifying remote sensing imagery is illustrated in (Figure 4b) where clouds over the front range in Colorado are misclassified as water pixels. In addition, because the Landsat ‘after’ image was collected on 17 September, 7 days after the beginning of the flood event, it is likely that the maximum flood extent is not captured in this scene.

![Figure 4a](image1.png)

(a) Classification of water in Landsat 8 image collected 5 May, 2013.

![Figure 4b](image2.png)

(b) Classification of water in Landsat 8 image collected 17 September, 2013.

Figure 4: Water pixel classification using Landsat 8 data collected 5 May, 2013 (a) and 17 September, 2013 (b).

Following the collection of remote sensing data, contributed data were also collected, geolocated, and interpolated following the methods discussed in Section 3.3.2. The interpolated pixels were then overlayed on the remote sensing classification to give an enhanced indication of flood activity in the Boulder area (Figure 5). The use of supplemental data sources, such as the CAP aerial photos,
shows flooding in areas that was not captured by the satellite remote sensing. In the case of the Boulder flooding, the cloud cover over the front range and western parts of the City of Boulder, made the identification of water from satellite platforms difficult. The ability of planes to fly below cloud cover as well as to collect data without the revisit limitations common to space-borne platforms, allowed the CAP to capture flooding and damage not visible from satellite images in the western parts of Boulder County (Figure 5, green pixels).

Furthermore, photos collected from UAVs flown over the Longmont and Lyons, CO areas, provided flood and damage information when manned aerial vehicles were grounded because of weather. At a finer scale, a change detection analysis using high resolution Digital Globe archive data as ‘before’ images (Figure 6 a,b) and the Falcon UAV photos as ‘after’ images (Figure 6 c,d) illustrates how aerial data collected from a UAV provide valuable, timely, high resolution details of flooding and affected transportation infrastructure (Figure 6 e,f). The use of these high resolution images allows flooding and damage to be identified within a few meters in near real-time.

4.2. Transportation Classification

Although the identification of water in the near-IR is a standard technique, the supervised classification of the Landsat data did not indicate any pixels as ‘water’ in the City of Boulder (Figure 4b). This could be because the image was collected on 17 September, a week after the flood event began, and flood waters could have receded, as well as the presence of obstructing vegetation and cloud cover. In addition, the high resolution Worldview-2 images, collected on 13, 14, and 17 September, also did not indicate the presence of flooding in the downtown area (Figure 7a). However, it is interesting to note that contributed data such as photos and tweets do indicate the presence of flooding in the City of Boulder. Using the geolocated tweets (n=130) containing the hashtag ‘#boulderflood’ geolocated near Boulder (105.0814° – 105.2972° W longitude and 40.09947° – 39.95343° N latitude) as well as geolocated photos (n=80), flood activity is indicated by local citizens in the downtown area (Figure 7b). While there may be uncertainties associated with information obtained from tweets, the presence of flooding and damage are more easily verified in photos (Figure 8).

Using contributed data points (Figure 9a), a flooding and damage surface is interpolated using a kernel density smoothing application as discussed in Section 3.3.2 for the downtown Boulder area. After an interpolated surface is created from each data set (tweets and photos), they are combined using a weighted sum overlay approach. The tweets layer is assigned a weight of 1 and the photos layer a weight of 2. A higher weight is assigned to the photos layer because information can be more easily verified in photos, therefore there is a higher level of confidence in this data set. The weighted layers are summed in ArcGIS 10.1, yielding a flooding and damage assessment surface created solely from contributed data (Figure 9b).
Figure 6: Water pixel classification in Falcon UAV image
(a) Classification of water in Worldview-2 images from 13, 14, and 17 September 2013. Although water is classified in Boulder County and the surrounding counties, there is very little flooding indicated in the City of Boulder.

(b) Classification of water in Worldview-2 images from 13, 14, and 17 September 2013 and geolocated contributed photos and tweets.

Figure 7: Worldview-2 remote sensing data does not capture flooded pixels in the City of Boulder (a), while contributed data indicate the presence of flooding (b).
surface is then paired with a high resolution road network layer. Roads are identified as potentially compromised or impassable based on the underlying damage assessment (Figure 9c). In a final step, the classified roads are compared to roads closed by the Boulder Emergency Operations Center (EOC) from 11 - 15 September 2013 (Figure 9d).

5. Conclusions

This paper presents a new methodology for locating natural hazards using contributed data, in particular Twitter, and the tasking of remote sensing data for these areas. Once remote sensing data are collected, they, and in combination with contributed data, can be used to provide an assessment of the ensuing damage. While Twitter is effective at identifying ‘hot spots’ at the city level, at the street level other sources provide a supplemental source of information with a finer detail (e.g. photos). In addition, remote sensing data may be limited by revisit times or cloud cover, so contributed ground data provide an additional source of information. Challenges associated with utilizing contributed data, such as questions related to producer anonymity and geolocation accuracy as well as differing levels in data confidence make the application of these data during hazard events especially challenging. In addition to identifying a particular hazard, in this case flood waters, by pairing the interpolated damage assessment surface with a road network creates a classified ‘road hazards map’ which can be used to triage and optimize site inspections or task for additional data collection.

The application of this methodology is continuous where data are constantly being collected during the progression of the hazard event (Figure 10). This will aid in refining the extent of the hazard, illustrate its progression over time, as well as assist in predicting new areas which may become affected.

6. Acknowledgments

Work performed under this project has been partially supported by the Office of the Assistant Secretary for Research and Technology, US Department of Transportation award # RITARS-12-H-GMU (GMU #202717). DISCLAIMER: The views, opinions, findings and conclusions reflected in this presentation are the responsibility of the authors only and do not represent the official policy or position of the USDOT/OST-R, or any State or other entity. Work performed under this project has been also partially funded by the Office of Naval Research (ONR) award #N00014-14-1-0208 (PSU #171570).

References


Figure 9: Using contributed data geolocated in the downtown Boulder area (a), an interpolated damage surface is created (b) and when paired with a road network, classifies potentially compromised roads (c). Roads which were closed by the Boulder Emergency Operations Center (EOC) that were also classified using this approach (d).


Tate, C., Frazier, T., 2013. A GIS Methodology to Assess Exposure
Agent-based Large-Scale Emergency Evacuation Using Real-Time Open Government Data

Wei Lu, Ph.D.
Phone: 865-576-9297
Email: luw4@ornl.gov

Cheng Liu, Ph.D.
Phone: 865-576-5593
Email: liuc@ornl.gov

Budhendra L. Bhaduri, Ph.D.
Phone: 865-241-9272
Email: bhaduribl@ornl.gov

Geographic Information Science and Technology group
Computational Science and Engineering Division
Oak Ridge National Laboratory
PO BOX 2008 MS6017
Oak Ridge, TN 37831-6017
Abstract

The open government initiatives have provided tremendous data resources for the transportation system and emergency services in urban areas. This paper proposes a traffic simulation framework using high temporal resolution demographic data and real-time open government data for evacuation planning and operation. A comparison study using real-world data in Seattle, Washington is conducted to evaluate the framework accuracy and evacuation efficiency. The successful simulations of selected area prove the concept to take advantage of open government data, open source data, and high resolution demographic data in emergency management domain. There are two aspects of parameters considered in this study: user equilibrium (UE) conditions of traffic assignment model (simple Non-UE vs. iterative UE) and data temporal resolution (Daytime vs. Nighttime). Evacuation arrival rate, average travel time, and computation time are adopted as Measure of Effectiveness (MOE) for evacuation performance analysis. The temporal resolution of demographic data has significant impacts on urban transportation dynamics during evacuation scenarios. Better evacuation performance estimation can be approached by integrating both Non-UE and UE scenarios. The new framework shows flexibility in implementing different evacuation strategies and accuracy in evacuation performance. The use of this framework can be explored to day-to-day traffic assignment to support daily traffic operations.

Keywords: open government data, urban transportation systems, high resolution data, evacuation management,
1. Introduction

Global urbanization has boosted the population growth unprecedentedly, which has also increased the travel demand and transportation infrastructure demand in urban areas. Natural or man-made disasters (e.g. Atlanta ice snow in 2014, Boston Marathon bombings in 2013) have tremendous impacts on urban transportation systems. How to maintain essential transportation operations during critical infrastructure disruption is an important issue.

The increasing availability of open government data has provided many new opportunities. Robinson et al. explained the benefits of using government data in public services (Robinson, Yu, Zeller, & Felten, 2008). Ding et al. introduced an open source portal to support the deployment of linked open government data in today’s international open government initiatives (Ding et al., 2011). Janssen discussed current policies and practices of using public sector data in Europe. He also proposed the legislation concerns about the conflict between data access and re-use (Janssen, 2011). Since the data.gov became available online in 2009, many applications were developed to provide useful information for publics (Hendler, Holm, Musialek, & Thomas, 2012). However, most of these researches and applications are not used in emergency evacuation areas, especially for real-time operations. To meet the Obama open government initiatives, Seattle government provides real time fire 911 calls data publicly, which updates every five minutes (data.seattle.gov, 2014), which can be used in real-time emergency evacuation planning effectively. This provides the opportunity to discover urban population dynamics during emergency situations.

Agent-based traffic simulation models have been well explored in traffic and transportation planning and operation domain. Agent-based behavior models are developed to describe the individual driving behaviors from system level to intersection level (Dia, 2002; Doniec, Mandiau, Piechowiak, & Espié, 2008; Nagel & Flötteröd, 2009). Besides these agent-based modeling studies for transportation simulation in relative small areas, large-scale agent-based microscopic traffic
simulation based on queuing theory is explored for multiple applications and empirical case studies in Europe (Balmer, Axhausen, & Nagel, 2006; Balmer, Cetin, Nagel, & Raney, 2004; Cetin, Burri, & Nagel, 2003; Meister et al., 2010). Evacuation planning and operation also needs simulation-based studies while real world evacuation data is hardly available. Microscopic simulation package VISSIM and macroscopic simulation package DynaSMART-P are compared for emergency evacuation scenarios and emphasized the advantage of using intelligent transportation systems to route evacuees (Han, Yuan, Chin, & Hwang, 2006; Yuan & Han, 2009). Chen and Zhan (2006) used an agent-based technique to model traffic flows at individual vehicle level and revealed that the road network structure and population density have impacts on evacuation strategies. Jha, Moore, and Pashaie (2004) used microscopic simulation model (MITSIM) to model the evacuation of Los Alamos National Laboratory. Cova and Johnson (2002) presented a method to develop neighborhood evacuation planning with microscopic traffic simulation in the urban - wildland interface. TRANSIMS (Transportation Analysis and Simulation System) was also used in evacuation simulations. Lu and Han discussed the impacts of zoning resolutions on evacuation performance with various travelers’ compliance rates (Lu, Han, Liu, Tuttle, & Bhaduri, 2014). Naghawi and Wolshon (2011) discussed multi-modal evacuation network performance through microscopic traffic simulation. Agent-based simulations can provide detailed information of individual vehicles, which helps evacuation operation managers to produce detailed evacuation strategies.

In order to take advantage of the real time open government data, especially under abnormal conditions, we propose a computational framework, World Wide Emergency Evacuation (WWEE), which enables emergency managers to use open data and detailed simulations in a scalable, real-time manner for efficient traffic operations. A comparison study using Seattle fire 911 data for evacuation simulation is conducted to evaluate the framework accuracy and evacuation efficiency. Two aspects are considered, including user equilibrium condition of
dynamic traffic assignment models (User Equilibrium vs. Non User Equilibrium), and temporal population resolutions in evacuation start times (daytime vs. Nighttime). The simulation results indicate that the effectiveness of our WWEE framework and various impacts of data and model resolutions on evacuation performance. Subsequent discussions on improving evacuation simulation accuracy and efficiency through high resolution demographic data are also presented.

2. Framework Description

LandScan population data and open-source agent-based traffic simulation package TRANSIMS provide valuable data and programs to build an efficient and accurate evacuation assignment and simulation platform. LPC-based traffic assignment brings many new issues to integrate modules in TRANSIMS package. A new evacuation framework built on several programs in TRANSIMS is proposed, as shown in Figure 1.

There are nine modules in this agent-based evacuation modeling framework, which convert the raw input data at the beginning to detailed simulation performance results at the end. It can provide both aggregated traffic information and individual vehicle tracking analysis.

1. Selection. According to the Seattle Fire 911 records, select the longitude and latitude values from Seattle open data website as the centroid of fire evacuation zone. The original input data are LandScan USA population (daytime/nighttime) and OpenStreetMap (Haklay & Weber, 2008) road network data. The users define a evacuation area by circle or polygon based on the selected centroid point. All the affected population and origins/shelters locations are summarized in two new files, as selected population and evacuation area. LandScan Global population data and OpenStreetMap (OSM) can be implemented for simulating areas outside US.
Fig. 1. Flow Chart of World-Wide Emergency Evacuation framework
2. Conversion. To take advantage of TRANSIMS simulation tools, the selected OSM road network is converted to TRANSIMS-based format. The population is converted to vehicle agents based on demographic information.

3. Mapping. The mapping module assigns the travelers in each LandScan cell to their nearest activity locations. Vehicles access to the network through activity locations along the road. This module is used to resolve the limitation of activity-based traffic simulation in TRANSIMS. The original program assumes that there are at least two activity locations in a Traffic Analysis Zone (TAZ) for inner-zone trips and at least one activity location for inter-zone trips. This is understandable in TAZ since the TAZ is usually large enough to include several road segments. But LandScan Population Cell (LPC) size is much smaller than TAZ and there are many cells having no roads crossing through their areas. In this case, the trips in that zone cannot be assigned to the activity locations. Also, the original TRANSIMS algorithm evenly assigns all the trips in one TAZ to all the activity locations in that zone. In the real world, travelers prefer to access roads at the nearest spots to their locations.

4. Distribution. Each LandScan USA population cell (LPC) is assigned as one origin zone. All the exit nodes generated in the first module are assigned as destinations/shelters. The OD matrix can be generated as formula (1) with minimum travel cost to each LPC. The travel cost can be calculated as equation (2), where impedance can be summarized with different shortest-path based routing algorithms, including shortest network distance routing, highway-based routing, and straight-line distance routing. In this evacuation model, only the shelters out of the evacuation area are considered.

\[
\text{Min} \{\text{Travel Cost}_i, \ i \text{ is the } i^{th} \text{ LPC} \} \tag{1}
\]

\[
\text{Travel Cost}_i = \sum_j \text{Impedance}_{ij}, j \text{ is the } j^{th} \text{ link in a route} \tag{2}
\]

5. ConvertTrips. This is a program provided by TRANSIMS to generate travel plans for every individual traveler with OD-matrix and used-defined departure
time choice model (or loading curve). Trip chain can also be implemented. But in this paper, we assume that all the vehicles travel to shelters directly after evacuation order is placed. TRANSIMS only consider 1000 TAZ zones as maximum for trip assignment. There are more than 1000 LPC cells in Seattle. Technically, each LPC is equivalent to one TAZ zone. This limitation has been adjusted to accommodate LPC-based traffic assignment method.

6. Routing Stabilization. To save the computational time, link-based trip assignment is implemented to generate macroscopic system performance results. There is a loop to optimize the results to achieve stable status. Volume/Capacity ratio, travel time change ratio, et al. can be used as the condition statement. After the macroscopic simulation is stable, the results can be pulled to the last visualization module for analysis.

7. Microsimulation Stabilization. Microscopic simulation is also implemented if the users want to see detailed simulation results of each vehicle for operation management. Similar to the module 6, users can use those ratios to decide the number of iterations to achieve stable condition. The simulation results are also able to be plotted for intermediate analysis.

8. Dynamic UE Convergence. To improve the simulation accuracy, dynamic traffic assignment is also integrated to simulate user optimized evacuation dynamically. The convergence conditions are defined by users, such as less than 2% changes in travel time for each vehicle. The final outputs with both system level information and individual vehicle movement are summarized.

9. Visualization. Beyond the research analysis with output data, the simulation results file is processed to display on a web-based application front-end with both macroscopic and microscopic information.

This framework provides a one-stop application tool for both researchers and practitioners. To test our concept of using open government data, we implement a case study in Seattle. However, our framework can also handle simulations outside US with LandScan Global data and OSM networks.
3. Evacuation Case Study Design

To assess evacuation efficiency with this proposed WWEE framework in different scenarios, an evacuation case study using data in Seattle, Washington is conducted through comparing with conventional TAZ-based traffic assignment.

3.1. Data resources

The Seattle real time fire 911 calls data provides real-time emergency situations monitoring within the city. It is updated every five minutes. The attributes include address, type, date/time, location (latitude, longitude), and incident number. Users can choose the location by different types, which contains aid response, medic response, auto fire alarm, auto medical alarm, bark fire, etc. In practice, users can analyze a group of records and define a centroid point of affected area. In our case study, we just choose one location with fire alarm for demonstration purpose. We use the latitude/longitude values of a record near downtown Seattle area as the centroid to draw a elliptical shape (as shown in Figure 2), which is equal to a circle area after projection to Cartesian coordinates system used in traffic simulation.

As stated in the WWEE framework, population distribution and road networks are two major input data for the evacuation assignment. The road network of selected Seattle area is shown as black lines in Figure 2. The background map source is from OpenStreetMap in Esri ArcGIS 10.1 software package.

LandScan USA population cell (LPC) data has a high resolution population distribution to provide 90m x 90m (3’×3’) resolution with national population distribution data (Bhaduri, Bright, Coleman, & Urban, 2007). It is much more accurate than conventional TAZ because some TAZ zones are large in scale or dense in population (Lu, 2013; You, Nedović- Budić, & Kim, 1998). Compared to TAZ-based traffic assignment that generates trips from large-scale zones, LPC-based methods allow small-scale, cell-to-cell trip generation.
In addition, LPC provides both day-time and night-time population distributions. The daytime is defined from 6:00am to 6:00pm and the nighttime is the left 12 hours. The LPC daytime data consists of 6614 non-zero cells, as shown in Figure 3a. The color represents cells with the number of population gradually, from red as high dense population to yellow as low population. Technically, TAZ and LPC have the same zone definition, but LPC size is much smaller. The total daytime population in Seattle is 395,670 with a cell of maximum population as 6059. The LPC nighttime data is consisted of 4851 non-zero cells as displayed in Figure 3b. The total nighttime population is 159,363 with a cell of maximum population as 948. During the daytime, the total population is almost doubled compared to nighttime situation.
Fig. 3a. LandScan USA 2011 daytime population data in selected Seattle area

3b. LandScan USA 2011 nighttime population data in selected Seattle area
3.2. Traffic Simulation Modeling

Based on the selected Seattle road network map in Figure 2, a detailed network configuration is generated for the agent-based microscopic simulation. This includes 2136 nodes, 2336 links, 1614 activity locations, and 392 traffic controls. The Seattle area is about 42.2 sq. mi. In our case study, only personal vehicle mode is considered. The traffic mode, such as transit, can be implemented later. The trip chain for each traveler is “walk-drive-walk”, which means walking to the car from origin, driving to the nearest parking lot at destination, and then walking to assigned shelters. After considering the vehicle per capita ratio and car pool possibility during evacuation in Seattle, the total trips for evacuation in selected area are 168,520 in daytime and 106,150 in nighttime as travel demand. In addition, 45 shelters are assigned and connected one-to-one with 45 exits of the road network. Evacuation trips generally move outward as evacuees leave an evacuation area and seek safety.

For the departure time choice models, S-shape loading curves are most accepted with empirical studies (Murray-Tuite & Wolshon, 2013). We adapted the Weibull distribution as our default model in WWEE. The Weibull distribution model is given by

\[ D(t) = 1 - \exp \left( -\beta t^\gamma \right) \]  

(1)

where \( D(t) \) is the cumulative percentage of people who left at time \( t \). The values of \( \beta \) and \( \gamma \) determine the shape of the distribution for faster or lower response to evacuation. We defined most people tried to access the road within the first five hours after evacuation orders placed.

Trip distribution models explain where people are willing to go as their safe destinations. In WWEE, we assume travelers will choose their nearest shelters outside the evacuation area. The evacuees in each LPC cell as \( S \) choose the nearest destination out of \( D \) based on travel time. The objective function is to find the minimal initial travel time for people in each LPC cell. If using Dijkstra’s shortest
path algorithm with forward star structure, we have to run S times shortest path algorithm to find the O-D pair for each source node and destination node, where S is the number of LPC source nodes. The time complexity is $O((E + V) \cdot \log V)$. If we connect all the destinations to a dummy super node and reverse the network with backward star structure, as shown in Figure 3, we can have the same O-D matrix output but with just one time run of Dijkstra’s algorithm. The time complexity for our proposed algorithm is $O(((E + D) + (V + 1)) \cdot \log(V + 1))$. It saves the computational time significantly. After testing on five different sizes networks, our algorithm reduced the computational time exponentially (500 to 45,000 times faster than conventional Dijkstra’s shortest path algorithm).

![Demonstration of super node trip generation](image)

**Fig. 4.** Demonstration of super node trip generation

Traffic assignment models determine how travelers get access to the road network and which routes they choose to approach destinations. As stated earlier, we use activity based traffic assignment model to connect each LPC to its nearest activity location. User equilibrium is commonly used for normal traffic simulation to achieve better representation of real-world traffic scenarios through multiple iterations. However, there is a controversial debate about the effectiveness of user equilibrium in evacuation simulations (Pel, Bliemer, & Hoogendoorn, 2012). The traffic pattern and travelers’ behavior are quite different during emergency evacuation scenarios. Travelers might only use the most used roads as their
evacuation routes, which is also one reason for highway congestion during empirical observations of historical evacuation events. Here, we implemented dynamic traffic assignment with and without user equilibrium (UE) conditions and let users to choose their best models. Meanwhile, multiple-threads based parallel traffic simulation model based on cellular automata theories is also completed to provide comparison between macroscopic and microscopic simulations.

3.3. Simulation Scenarios

There are two aspects of parameters considered in this study: traffic assignment models resolution (simple Non-UE vs. iterative UE) and data temporal resolution (Daytime vs. Nighttime). In total, 4 evacuation scenarios are modeled in this paper: NUE-Daytime, UE-Daytime, NUE-Nighttime, and UE-Nighttime. The proposed framework was run on a Window 7 64-bit laptop computer. The configuration of this laptop is 16GB RAM, 2.6GHz Intel(R) Core(TM) i5-3320 CPU with 4 cores, and 500GB hard disk. A 12-hour simulation time is used and the other parameters are set as stated in the preceding sections. To adjust the impact of random number in the framework, all the simulation results are based on the average of 30 independent runs. To evaluate the evacuation efficiency of microscopic simulations, we used the evacuation clearance time, evacuee arrival rate, average individual travel time, and model computational time in this study (Han, Yuan, & Urbanik II, 2007).

4. Simulation Results and Discussions

Evacuation simulation results from comparison studies of the selected Seattle area are summarized in Table 1. The non user equilibrium (NUE) assignment in daytime scenario needs the longest evacuation clearance time. First, the daytimes trips is about 1.5 times as the nighttime trips, which increases the system total travel time and congestion possibility. Second, the NUE assignment doesn’t consider the situation that travelers might use alternative routes to approach their destinations. In this situation, some road sections (such as highway and major city
roads) are highly used when travelers don’t have the information about other roads conditions. Thus, for the same temporal resolution, daytime or nighttime, NUE scenarios need more time to clear evacuation than UE. For the same traffic assignment condition in Seattle case study, evacuation during the daytime needs more time to achieve clearance than nighttime. The average travel time follows the similar patterns as evacuation clearance time. The longer clearance time is mainly caused by congestion on the road network during evacuation. The congestion increases travelers’ average travel time. The computation time is the time period for the computer to finish one scenario simulation. In either daytime or nighttime scenarios, UE needs more computation time than NUE. To achieve user equilibrium condition, iterative simulation is run until reaching the UE status. There are 7 iterations for the daytime scenario and 2 iterations for the nighttime scenario. Basically, the iteration helps reduce the congestion conditions by re-routing some travelers to alternative routes.

Table 1. Evacuation simulation summary for different scenarios

<table>
<thead>
<tr>
<th>Temporal Resolution</th>
<th>Traffic Assignment Conditions</th>
<th>Evacuation Clearance Time (hours)</th>
<th>Average Travel Time (minutes)</th>
<th>Computation Time (minutes)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daytime</td>
<td>NUE</td>
<td>14.8</td>
<td>38.3</td>
<td>16.3</td>
</tr>
<tr>
<td></td>
<td>UE</td>
<td>11.9</td>
<td>26.9</td>
<td>72.2</td>
</tr>
<tr>
<td>Nighttime</td>
<td>NUE</td>
<td>7.9</td>
<td>19.3</td>
<td>8.1</td>
</tr>
<tr>
<td></td>
<td>UE</td>
<td>6.5</td>
<td>17.4</td>
<td>15.8</td>
</tr>
</tbody>
</table>

The impacts of temporal high resolution data with daytime and nighttime formats under user equilibrium condition are summarized in Figure 5. Similar in table 1, the daytime scenario needs more time to achieve 100% arrival rate. Under the same travel demand model, fewer trips in nighttime scenario improve evacuation efficiency than daytime scenario. However, during real-world emergency evacuation, we cannot expect fewer trips during the daytime.
Emergency managers and transportation planners can make efficient policy to control the travel demand, such as staged evacuation strategy. It gives the priority to people in seriously emergent area to access the road network for evacuation. The LandScan high resolution population data can also help estimate the critical areas to maximize the evacuated population.

![Graph](image)

**Fig. 5.** Evacuation arrival rates for daytime and nighttime comparison study

Figure 6 illustrates the impact of user equilibrium conditions of traffic assignment on evacuation performance in both daytime and nighttime scenarios. For the 12 hours simulation in daytime scenarios, the UE assignment achieves 100% arrival rates, but the NUE assignment only reaches 89.1% arrival rates. The cases in nighttime scenarios have similar pattern with less simulation time to achieve 100% arrival rates. In these four scenarios, UE assignment improves evacuation efficiency by re-routing trips to alternative routes. However, travelers might not have the real-time system road condition information during emergency evacuation. They would still use the pre-defined routes after they access the road network if there is no further information provided. Real time traffic information
through intelligent transportation systems can expand travelers’ knowledge about road conditions and overall performance. This helps travelers to re-route during evacuation to find the shortest path to arrive destinations. For a practical estimation of evacuation performance for planning purpose, the parameters values about evacuation efficiency locate between UE situation and NUE situation.

**Fig. 6.** Evacuation arrival rates for assignment condition comparison study

In order to improve user experience for various clients, we developed a visualization tool vehicle-based microscopic analysis. The web-based microscopic visualization tool provides detailed animation of vehicle movement, which helps identifying the network choke points during evacuation. The visualization tool can also help users to re-run their models or to test different evacuation strategies. Figure 7 shows a snapshot of microscopic visualization. Each point means a vehicle. Users can select different simulation areas from the website http://gisthal.ornl.gov/wwee. The animation time step ranges from 1 second to 15 minutes. The 15-minutes interval animation produces equivalent outputs of the macroscopic visualization. Clients can select certain time to indicate using LandScan daytime or nighttime population data for evacuation simulations. Clients can also drag the control button to display results at any time slot during an evacuation simulation.
5. Conclusions

In this paper, a World-Wide Emergency Evacuation (WWEE) framework using high resolution data, open source data, and open government data is presented for evacuation planning and simulation. To test the effectiveness and efficiency of this framework, comparison studies using population data in different temporal resolutions and traffic assignment conditions are conducted. Through simulation studies of those four scenarios in selected Seattle area, three major findings are concluded here.

- Open government data and open source data provide great opportunity to explore urban transportation systems under emergency evacuation scenarios. There is still space to improve the open government data quality for wider usage. For example, the traffic count data from Seattle Department of
Transportation is out-dated. It only provides 2007, 2008, 2009 years data. This kind of data can be used to validate and calibrate traffic simulation models.

- Urban population dynamic information varies according to temporal resolutions. In this Seattle case study, the daytime scenario requires more evacuation preparation time compared to nighttime scenarios. Most existing studies using Census data are actually only consider the nighttime scenarios. The temporal data resolution can improve model accuracy.
- Traffic assignment with user equilibrium condition is commonly used for traffic simulation modeling. But it does not consider the real emergency situation well. A better estimation can be reached by integrating both UE and non-UE scenarios.

This high spatiotemporal resolution traffic assignment framework can be easily expanded to simulate any area in the world since LandScan Global has the whole world demographic data and OpenStreetMap provides the worldwide road network for free. The potential applications of this framework are not limited to evacuation studies. With daily demand-modeling framework (Balmer et al., 2006), the normal daily traffic operation and prediction can be conducted to provide various application purposes for researchers and practitioners. Though implementation of intelligent transportation systems with modern communication technologies, such as connected vehicle technology (Lu, Han, & Cherry, 2013) and high performance computing (Karthik, 2014), to provide real-time travel information, a more accurate microscopic dynamic traffic simulation can be developed for various application purposes.

**Acknowledgments**

This manuscript has been authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-
exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The authors would like to acknowledge the financial support for this research from the US Government for the development of LandScan USA model and database.

References


ABSTRACT:

Critical infrastructure disruption, caused by severe weather events, natural disasters, terrorist attacks, etc., has significant impacts on urban transportation systems. We built a computational framework to simulate urban transportation systems under critical infrastructure disruption in order to aid real-time emergency evacuation. This framework will use large scale datasets to provide a scalable tool for emergency planning and management. Our framework, World-Wide Emergency Evacuation (WWEE), integrates population distribution and urban infrastructure networks to model travel demand in emergency situations at global level. Also, a computational model of agent-based traffic simulation is used to provide an optimal evacuation plan for traffic operation purpose [1, 58]. In addition, our framework provides a web-based high resolution visualization tool for emergency evacuation modelers and practitioners. We have successfully tested our framework with scenarios in both United States (Alexandria, VA) and Europe (Berlin, Germany) [2]. However, there are still some major drawbacks for scaling this framework to handle big data workloads in real time.

On our back-end, lack of proper infrastructure limits us in ability to process large amounts of data, run the simulation efficiently and quickly, and provide fast retrieval and serving of data. On the front-end, the visualization performance of microscopic evacuation results is still not efficient enough due to high volume data communication between server and client. We are addressing these drawbacks by using cloud computing and next-generation web technologies,
namely Node.js, NoSQL, WebGL, Open Layers 3 and HTML5 technologies. We will describe briefly about each one and how we are using and leveraging these technologies to provide an efficient tool for emergency management organizations.

Our early experimentation demonstrates that using above technologies is a promising approach to build a scalable and high performance urban emergency evacuation framework that can improve traffic mobility and safety under critical infrastructure disruption in today’s socially connected world.

**KEYWORDS:** emergency evacuation framework, scalability, cloud computing, Node.js, WebGL, HTML5.

1. **INTRODUCTION:**

Natural or man-made disasters, such as Atlanta ice snow in 2014 and Boston Marathon bombings in 2013 have tremendous impacts on urban transportation systems. It is critical for emergency managers and transportation professionals to have an efficient evacuation plan during these kinds of infrastructure interruptions. Simulation-based studies are commonly used in evacuation planning and decision making processes. Most existing simulation models in transportation evacuation are restricted to certain geographic areas and selected traffic simulation tools, such as OREMS, DYNASMART, VISSIM, etc. [3]. Each tool has its own data format requirement and simulation preferences (macroscopic, mesoscopic, or microscopic). This leads to the need of a uniform simulation tool in evacuation planning community, so that researchers can focus on improving traffic models in travel demand, trip distribution, and traffic assignment areas.

This motivated us to develop a computational framework and easy to use system, called World-Wide Emergency Evacuation (WWEE). It is a rapid response emergency evacuation modelling system that estimates evacuation time and travel conditions (Speed, Congestion, etc.) for any geographic locale in the world. WWEE integrates population distribution and urban
infrastructure networks to model travel demand in emergency situations at global level. Also, a computational model of agent-based traffic simulation is used to provide an optimal evacuation plan for traffic operation purpose [1, 58]. In addition, our framework provides a web-based high resolution visualization tool for emergency evacuation modelers and practitioners.

However, there are still some major drawbacks for scaling this framework to handle big data workloads in real time. On our back-end, lack of proper infrastructure limits us in ability to process large amounts of data, run the simulation efficiently and quickly, and provide fast retrieval and serving of data. On the front-end, the visualization performance of microscopic evacuation results is still not efficient enough due to high volume data communication between server and client.

We are addressing drawbacks on back-end by using cloud infrastructure, high performance NoSQL database for storage, and Node.js to solve I/O bottlenecks. On the front-end, to provide rich and blazingly fast interactive visualization, we use next-generation web technologies namely WebGL, Open Layers 3 and HTML5 collection of technologies.

The rest of the paper is organized as follows. Section 2 describes background and motivation for the framework. In Section 3, we describe our system architecture and in detail, various components and technologies we are using and leveraging to scale our framework. Section 4 presents the conclusion of the paper and our future work is described in Section 5.

2. BACKGROUND AND MOTIVATION

Simulation-based evacuation studies give researchers and practitioners great tools to evaluate evacuation strategies. A detailed review of various emergency evacuation models by Alsnih et al. [4] pointed out the three main procedures to devise emergency evacuation plans, including evacuees’ behavior analysis, transportation engineering analysis, and the role of government. The interaction and cooperation among these three aspects are needed to provide better solutions of a mass evacuation on the current transport network. Various scales of evacuation areas were conducted to evaluate the evacuation efficiency and systems performance, from the whole state Tennessee to a corridor in Washington D.C. [5, 6]. Microscopic traffic simulation is becoming
more popular than conventional macroscopic traffic simulation in evacuation study due to the cheap but fast computing capacity and expectation of detailed performance. Jha, et al. [7] took advantage of microscopic simulation model (MITSIM) to model the evacuation of a small area - Los Alamos National Laboratory. Cova, et al. [8] presented a method to develop neighborhood evacuation planning with microscopic traffic simulation in the urban – wildland interface. Household-level evacuation planning is implemented in various scenarios. Activity-based traffic simulation provides more realistic simulation at the traffic assignment stage, which is adopted by a population simulation package called Transportation Analysis and Simulation System (TRANSIMS) [9]. Henson et al. [10] reviewed 46 activity-based models and used TRANSIMS to demonstrate their competency for homeland security applications. Despite these existing research efforts in evacuation simulations, most of these studies assume a one-to-one trip assignment to evacuees’ full compliance for predetermined destination and route assignments. This motivated us to develop such a computational framework.

The goal of WWEE is to develop an easy-to-use system targeted at emergency response planners. WWEE consists of software and data developed in ORNL combined with readily available data from the open source domain. There are four major components in this system: 1) LandScan Global, a high-resolution population distribution database, developed at ORNL, and a public domain global street network, OpenStreetMap (OSM), 2) a web-based network editing tool that is used to extract the street network from OSM and refine network as needed and include traffic controllers, 3) an open source traffic simulation model (TRANSIMS) in which macroscopic modelling is performed as well as agent based microscopic simulation is performed, an open source time-based microscope traffic simulation model from MITSIM, and 4) a web-based user interface that can display either the agent-based micro individual vehicle analysis or link based macro traffic analysis results.

The user logs into the main page and defines an evacuation area using several drawing tools (circle, rectangle, polygon, etc.). The system will calculate the number of evacuees within this evacuation area from LandScan Global. The node-link based network is then extracted from OSM and converted it to a traffic-modeling network, which includes lanes, pocket, connectors, intersections, and controllers. If the extracted street network does not present as the real street network, the user can use the graphic network-editing tool to fine-tune the local network.
configuration or controllers. The user also has the option to choose one of three traffic simulation methods included (TRANSIMS macro, Router, TRANSIMS agent based, Simulator, or MITSIM micro). Results are displayed on a web-based user interface when the simulation is running.

3. SYSTEM ARCHITECTURE

We have already described about the drawbacks with our framework and our approach to address the same earlier. In this section, we will describe each of the components used in our system architecture. An illustration of our architecture is shown in Figure 1.

![Figure 1: Scalable System Architecture](image)
I. Cloud Infrastructure:

Cloud computing has gained enough awareness that majority of urban planning organizations and researchers know the benefits it can offer to an organization [11, 51]. Many organizations who were asking “Will moving to cloud solve our problems?” are asking “When can we move to the Cloud?” [11, 17]

With increasing adoption of Cloud computing, various infrastructure solutions have been built that ranges from managing nodes to Operating Systems (OS), software platform, applications built on top of them, to networks and storage. One such leading solution that provides these capabilities is called OpenStack.

**OpenStack** is a free and open source infrastructure solution that powers many of the clouds used in Fortune 500 companies [11]. It provides broad range of components for controlling compute, storage and networking resources needed for massively scaling and deploying resources in global data centers. Originally founded by Rackspace Hosting and NASA, it has grown to be a global active developer community [11, 12].

Due to this broad nature, deploying the plain vanilla distribution of OpenStack is painful and difficult, especially for smaller organizations [13]. It requires dedicated set of expertise and can make moving to cloud very time-consuming. Organizations are typically requiring a bigger bang for their buck, i.e. they want to have most of the features found in large-scale commercial cloud deployments, but with limited time or investment or both [13, 14].

**Ubuntu**, a leading Linux distribution has been bundling OpenStack with their service since 2011 [13]. Known for making use of Linux simpler on desktops and server, they have done the same with OpenStack. Building a large-scale enterprise cloud, from provisioning to installation, deployment and management has been simplified and made easy to use with Ubuntu “Metal as a Service” (MAAS) tool [15, 16, 17].

OpenStack and Ubuntu have a long history of tight and continuous integration and rigorous testing, making their combination one of the fastest and reliable ways to build a cloud. Additionally, Ubuntu provides a five year support of this combination, with their Long Term Support (LTS) releases [15, 16].
Ubuntu, the most popular OS and widely used developer OS for OpenStack, has been historically used for powering WWEE simulator [15, 16]. Ubuntu supports conventional and old hardware pretty well [17, 18]. Due to the above mentioned benefits, we decided to adopt Ubuntu OpenStack as our cloud infrastructure.

**Ubuntu MAAS:**

Ubuntu MAAS is a quick and easy-to-use tool for transforming machines into a cloud [17]. It makes it easy to set up hardware on which to run “Ubuntu’s OpenStack cloud infrastructure” [16, 17]. Machines (nodes) can be added, provisioned, commissioned, and deployed without interrupting rest of the network. Machines can be retired as well without any interruptions. It supports dynamically provisioning or re-allocation of physical resources to handle “Big Data” workloads [15, 16, 18].

Key components of MAAS tool are:

1. Region controller: consists of
   a. Web User Interface (UI) for easy and quick management of resources,
   b. REST API for integrating MAAS with third-party tools and workflows,
   c. Metadata server for cloud initiation, and
   d. DNS Server (Optional) [18, 19].
2. Cluster controller(s): controls provisioning and management of a cluster (or group of Nodes). It consists of
   a. TFTP (PXE) Server: booting Node(s) via network.
   b. DHCP Server (Optional): for dynamically assigning and managing IP addresses.
   c. Power management on Node(s): starting and stopping Node(s) [18, 19].
3. Nodes: Each node is an individual machine

DNS and DHCP are optional services because existing services used within an organization could be potentially used.

A typical MAAS deployment (as illustrated in Figure 2) has one Region controller, at least one Cluster controller, and Nodes needed to handle your workloads [18, 19]. Provisioning process is explained in detail in [19].
Experiment:

We will discuss the cloud infrastructure setup process with MAAS, challenges faced and solutions to handle the same. Our experiment consisted of 3 physical machines with configurations and their use in our cloud (as specified in Table 1.)

<table>
<thead>
<tr>
<th>Machine</th>
<th>Processors</th>
<th>RAM</th>
<th>Storage</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>gcloud2</td>
<td>Intel Core™ 2 Quad core CPU Q6700 – 2.66GHz</td>
<td>4 GB</td>
<td>430 GB</td>
<td>Region and Cluster Controller</td>
</tr>
<tr>
<td>gcloud3</td>
<td>Intel Core™ 2 Quad core CPU Q8400 – 2.66GHz</td>
<td>4 GB</td>
<td>320 GB</td>
<td>Node</td>
</tr>
<tr>
<td>gcloud1</td>
<td>Intel Xeon™ Octa core CPU X5460 – 3.20GHz</td>
<td>16 GB</td>
<td>3 TB</td>
<td>Node</td>
</tr>
</tbody>
</table>

Table 1: Machine (Node) configurations used for our experiment
Prerequisites: We started with installation of latest Ubuntu 14.04 MAAS distribution on one of the machines. The installations were always unsuccessful due to failing hard drives. We diagnoses the hard drives with SMART utility provided along with the hard disks and removed failing ones. As the other two machines were also old, we decided to check for hardware (memory and hard drive) failures before we proceeded with setting up cloud.

Region and Cluster Controller: We again started with installation of MAAS on gcloud2. The Region and Cluster controller was setup on the same machine. This is the ideal deployment strategy recommended for initial cloud setup [16, 18, 19]. Then, we completed post-install tasks like adding users and getting boot images. This process was smooth and easy due to Ubuntu’s commitment towards making the cloud infrastructure setup simpler. Also, the documentation by Ubuntu community helped as well.

DHCP and DNS Server: Next, we proceeded with setting up our own DHCP and DNS Servers. This was needed as to not conflict with our organization’s DHCP and DNS servers. All our nodes had to be on a private network, i.e. gcloud1 and gcloud3. Only our region controller, i.e. gcloud2 could be connected to our organization’s network. All the internet traffic from individual nodes had to be routed via gcloud2 via proxy.

We could not setup DHCP server initially via MAAS UI. So, we manually installed isc-dhcp – server, configured its properties and had it running as a separate DHCP server, disconnected with MAAS. Then, we shut down the separate DHCP server, replicated the same configurations in MAAS UI to have MAAS – DHCP server working together.

Proxy Configuration: We ran across an infamous problem with Ubuntu MAAS – If a private network is used, Proxy and DNS servers had to be manually configured with the private network addresses used. Detecting this problem was time-consuming. But, once we had figured it out, we will able resolve it quickly.

Provisioning: gcloud3 node was provisioned first and went through the following three steps:

Enlistment: When gcloud3 started, it obtained IP address from our DHCP server. Then, it was booted via PXE and retrieved Ubuntu boot image from the Cluster controller. The image ran an
initial discovery procedure to collect and send gcloud3’s configuration to the Cluster controller and registered itself in MAAS with a “Declared” state [19].

**Accept and Commission**: Once gcloud3 was enlisted, we commissioned the node. Base OS was installed on OS. But, it failed to get latest packages from internet. It was during this process, we realized this was due to above mentioned “Proxy Configuration” problem. Once we resolved the same, and re-ran the same process, it got and installed the latest packages from internet. Its state was changed to ‘Ready’, meaning it was ready for deployment [19].

**Deployment**: gcloud3 was allocated to one of our users.

We are conducting experiments on the best way to allocate our resources for various needs – one node for retrieving, processing and storing data; another node for communicating with simulation framework; and last node for managing our cloud infrastructure.

II. **Databases**

Our datasets consists of: 1) LandScan Global, a high-resolution population distribution database, developed at ORNL, and a public domain global street network, OpenStreetMap (OSM). Currently, these datasets are stored in PostgreSQL traditional relational database management system (RDBMS). The case study data of population and networks are preprocessed because PostgreSQL used in our framework cannot handle volume of datasets, primarily LandScan global population and OpenStreetMap network data.

NoSQL represents a radical paradigm shift designed to overcome certain limitations of RDBMS [20]. While RDBMS excels for applications that requires ACID transactions, it either provides very poor performance or fails in handling huge datasets with limited available hardware. Also, RDBMS is not well suited for real-time analytical needs [20, 21]. NoSQL solutions apply novel methods to not only overcome these limitations, but also provide a fast linear performance and are Cloud-capable as well [20, 21]. We are analyzing various types of NoSQL solutions to find the one best that can handle our data workloads and real-time analysis. Based on our preliminary analysis, we are inclining towards MongoDB.
III. Node.js

Node.js is a server-side JavaScript environment for easily developing scalable network applications. It’s built on top of Chrome’s V8 engine [22, 23]. Node.js uses an event-driven programming model, which is different from traditional multi-thread programming [24, 25, 26]. To easily understand comparisons between these models, let’s take the case of web servers as our example.

Traditional web servers require large amounts of I/O and multi-threading is an efficient way of using available processors (one thread per processor in a multicore system.) But, the downside is that cost of context switching between the threads is expensive, and thus limiting the speed and performance of the server [22, 26]. Event-based programming aims to provide an efficient alternative to address such drawbacks [26]. It utilizes a single event loop and uses asynchronous or non-blocking I/O model so execution do not get blocked when waiting for I/O operations [26, 28].

Node.js uses the event-based programming model to provide the power of multithreading, but with low overhead and a lock free design [28, 30]. It provides more control over the execution flow and makes it very attractive for developers as they do not have to handle the hard complexity of multi-thread programming mentioned above [30]. Various studies discussed in [28], [29] and [31] shows how Node.js outperforms traditional web servers such as Apache Web Server.

Node.js is increasing becoming popular for use on the server-side, due to its lightweight nature [22]. It’s ideal for (1) data-intensive real-time applications that run across distributed machines, (2) Websocket server like chat server, (3) Fast file upload client etc. [25].

For scaling WWEE, we are using JavaScript as one of the primary languages for both client and server-side. JavaScript has been traditionally used in client side, i.e. browsers for displaying webpages. Server-side applications require another programming language such as Java or PHP. But, Node.js uses JavaScript as its programming language. By using Node.js, we are writing both server and client side application with just one language and one codebase [22, 30].
On our server-side, we use JavaScript along with Node.js modules for retrieving and serving data quickly from the database; message and data passing to WWEE simulator as well as storing and caching the output data from the simulator; routing our web service requests to appropriate controllers; and administration of our application including logging. Our client-side visualization stack is built based on JavaScript libraries such as OpenLayers, ExtJS, Cesium, Three.js etc. We have detailed discussed about them and its use in our application below.

Also, we will discuss about Web Sockets, also belonging to HTML5 family of technologies, and how Node.js improves I/O performance on server side below.

IV. WebGL:

WebGL is a JavaScript API for rendering rich and interactive graphics including 3D in a browser without the need for plug-ins. It’s based on widely adopted OpenGL ES graphics library [33]. Full power of “Graphics Processing Unit” (GPU) accelerations can be harnessed for processing graphics computations and rendering them using just JavaScript and WebGL [34, 35]. Before WebGL, plug-ins or native applications were needed to be downloaded and installed to get a 3D experience [35, 36].

WebGL is one of the popular technologies within HTML5 family. Majority of browsers in desktop and mobile already support WebGL [35, 36].

WebGL is a low-level API and since it is increasingly being adopted, many libraries have been built on top of WebGL that provides various rich set of features and common utilities needed to produce high-end visualizations. While there are tons of libraries being developed, we will mainly discuss about three libraries used for our needs: (1) Open Layers 3, (2) Cesium, and (3) Three.js.

V. Open Layers 3 (OL3)

OL3 is a cutting edge and high-performance mapping library, being developed by OSGeo to provide rich user experience including 3D using HTML5, WebGL, Cesium, CSS3 and Closure
OL3 is re-written from scratch with a clean and modular design. But, at the same time, rich set of features and utilities available in Open Layers 2 are being ported to OL3 with the new design. OL3 supports mobile platforms and devices out of the box [45].

OL3 uses various techniques for creating high-performance maps such as better Garbage Collection (GC), re-using objects, uses rAF, avoiding unnecessary boxing or unboxing operations, and re-drawing as few pixels as possible [45]. Also, Google’s closure compiler is used for advanced optimizations such as code checking and dead code removal to produce high performance, lightweight and extremely compressed JavaScript file [44].

VI. Cesium

Cesium is a virtual globe and map engine based on WebGL. Built for precision and performance, it supports visualization of dynamic geospatial data and worldwide terrains [41, 42]. With just one API, three views – 2D, 2.5D Columbus and 3D Globe views can be created and switched easily. Cesium is free, open source and has an active community supporting it [46]. With WebGL and Cesium, OL3 is being built to support very large vector layers, 2D tilted and rotated views, and rich collection of 3D capabilities such as 3D globes creation and terrain in local projection [42, 43].

Three.js: Our application uses another WebGL based library called Three.js, as it supports various common utilities and helper functions such as textures, lighting, shaders etc. that makes use of WebGL simpler. [37].

VII. HTML5

HTML5 introduces (1) next version of HTML5 specification, and (2) Large collection of HTML5 Technologies (as illustrated in Figure 3) [32] used to enable rich and powerful applications. We will describe some of these technologies we use below.
Web Storage:

There are primarily three options in HTML5 for storing data on the client, i.e. browser: (1) Session Storage, (2) Local Storage, and (3) Indexed DB. We plan to use all of them for various purposes and details on them and its use are discussed below.

**Session Storage**: is a HTML5 technology used to solve the storage problems associated with cookies. Two main problems with the cookies are: (1) Separate instances of the web application cannot run, without interfering with each other typically, and (2) Storage limit is few kilobytes [38, 39]. Session Storage is fundamentally designed to solve both these problems. Data is stored in key / value pairs and its lifetime is limited to per tab or window only [38, 39]. We plan on using this storage for saving map and visualization states, including the currently selected area, visualization variables, zoom and resolution levels and device orientations.
Local Storage: is another technology for storing data in key / value pairs – but, data persists in the client beyond a page refresh [38]. As data is stored in client and does not need to be retrieved from the server after the first time, network traffic can be greatly reduced. Also, data is available for processing immediately. Majority of the browsers already support it [38, 39]. We plan to use this storage for storing visualization results and cached tiles.

But typically, Local Storage’s limit is capped around 5MB (depending on the browser) [38, 39]. So, as the size and number of our visualization results grows, we are forced to either discard or overwrite previous visualization data, or worse cannot use Local Storage at all. Hence, in the future, we plan to use another storage called Indexed DB that can solve this problem.

Indexed DB: is an upcoming API that provides a high-performance NoSQL-like data store for storing large amounts of structured data [40]. Indexed DB is also a low level API that allows you to create databases, data stores, indexes, populate and iterate data [39]. It can also support data revisions [39, 40]. Typically, there is no storage limit, making it suitable to store large datasets.

While this seems to solve our need for storing large amounts of data in the client side, this technology is very new. API is complex, difficult and likely subject for further revisions [39, 40]. Besides, older browsers and mobile devices do not support it [39]. Due to these drawbacks, it is too early to adopt this technology. But, we plan to use it in future, once it’s mature and stable.

CSS3 can be used to create rich animations. By moving the animations from scripts to styles, websites can be built with a cleaner codebase for visuals [42, 52]. Web Messaging and second version of XMLHttpRequest (AJAX) specifications support cross-origin communications. Additionally, standardizations for device orientations, notifications and touch events are in progress [42, 52]. We will use these upcoming technologies for rich visuals and supporting mobile devices once they are standardized.

Web Workers and Web Sockets, also part of HTML5 family of technologies are discussed in next sub-sections.
VIII. Web Workers

We have already discussed the problems of concurrency and complexity of multi-thread programming on server side earlier. This sub-section will discuss bringing threading to JavaScript on the client side.

Imagine a website that does common tasks - query and process large datasets, control DOM, and handle User Interface (UI) actions and events as well. As JavaScript is single threaded, so multiple JavaScript programs cannot run at the same time [47]. If they can be run concurrently, we can easily reduce the time needed.

Developers are already using few workarounds to handle this problem by using AJAX, setTimeout(), and event handlers [47]. Though these features are non-blocking and run asynchronously, they do not necessarily mean concurrency, i.e. only when one script has yielded, asynchronous events are processed [47, 48].

Web Workers in HTML5 aims to solve the problem, by bringing threading to JavaScript. Web Workers allows concurrent execution by breaking up huge tasks into small sets, so computationally intensive tasks such as large datasets processing and image filtering can be processed in the background in separate thread(s), without blocking the UI thread [47, 49].

To reduce typical concurrency problems such as deadlocks and starvation, Web Workers cannot access non-thread safe components or DOM [48]. The complexity of multi-thread programming and overhead to deal with common pitfalls such as deadlocks, starvation and race conditions makes it very hard [30, 31, 49]

However, there are few downsides of using Web Workers, due to memory cost and as it cannot access DOM as well. There are under active areas of research currently and resolutions to fix the same will be drafted in the near future [49, 50]. With JavaScript engines getting faster day by day, memory cost would come down. In future, we can even build faster websites, by running WebGL contexts in separate threads [49, 50].
IX. Web Sockets

The visualization performance of microscopic evacuation results is still not efficient enough due to high volume data communication between server and client. We have earlier discussed about our various approaches to solve this problem using: (1) Data storage on the client (browser) to store map tiles and simulation results, to reduce network traffic and improve performance, (2) Cloud Infrastructure, so dedicated nodes are available for sending data to client, and does not interfere with nodes used for simulation, and (3) Using fast read performance data storage system like NoSQL for retrieving and sending the data to the client.

There is another HTML5 technology called Web Sockets, which can improve our network performance as well. Typically, a client starts an HTTP connection to the server and keeps this connection alive until response has been sent. This is called “long polling” [52, 53]. But, the bottleneck is the overhead of HTTP, making it not well suited for low latency applications [53].

Web Sockets solves this problem by creating a persistent duplex connection between the server and the client [53]. Web Sockets enables server side applications to initiate and send data to the clients. Majority of the browsers already support it [52, 53].

Traditional web servers such as Apache Web Server use one thread per connection, and make them not well suited for large number of Web Socket connections [27, 53]. To support such large number of connections require a non-blocking architecture that provides high concurrency at low I/O cost - this architectural design is what is used in Node.js [27, 53].

Our Server Side application that is being built with Node.js, uses Web Sockets to send visualization data to the connected clients, instead of the clients constantly connecting, requesting data and overwhelming the server. With this approach, we hope to further save some network costs.

X. ExtJS

Last, but not the least, we will discuss about ExtJS, one of the leading JavaScript application frameworks, that is being used to create our interactive visualization web application. ExtJS 5
uses Model-View-ViewModel (MVVM) for building cleaner and modular codebase. It provides rich set of features – layout managers, data packages, two way data binding, delegated event model, highly optimized class system, and so on [41, 54]. Tons of plugins for quick creation of grids and trees, charting, and drawing are available [41]. ExtJS 5 is designed to support both mobile and web devices with one codebase [54]. Other notable benefits include dynamic loading, cross-browser compatibility, localization, unit and interface testing tools, debugging tools, good documentation, and active support community [41, 54].

4. CONCLUSIONS

In this paper, we have described various challenges and practices used to scale WWEE emergency evacuation framework. WWEE is used for simulations of urban transportation systems under critical infrastructure disruption in order to aid real-time emergency evacuation. On the back end, we have described need for an infrastructure to process large amounts of data, run the simulation efficiently and quickly, and provide fast retrieval and serving of data. We have addressed this drawback by using cloud infrastructure, high performance NoSQL database for storage, and Node.js to solve I/O bottlenecks. On the front-end, the visualization performance of microscopic evacuation results is still not efficient enough due to high volume data communication between server and client. We are using next generation web technologies such as WebGL, Open Layers 3 and HTML5 collection of technologies to provide rich and blazingly fast interactive visualization. Our early experimentation demonstrates that using above big data and web technologies is a promising approach to build a scalable and high performance urban emergency evacuation framework and an efficient tool for emergency management agencies to improve traffic mobility and safety in emergency evacuation scenarios.

5. FUTURE WORK

We have built the base cloud infrastructure using Ubuntu MAAS. We have discussed about various approaches to scale back-end and front-end system. Next, we would like to scale our
middleware, i.e. simulator using our cloud infrastructure and other cutting-edge technologies. We are currently analyzing the challenges involved in the same.

We plan to measure the performance of our front-end application using Stats.js - frames per seconds (FPS), milliseconds needed to render frame (MS) and allocated memory (MB) [55].

Our application use JSON to retrieve data from our servers, do post processing, and finally copy data to WebGL buffers. This approach is problematic for large-scale visualizations. WebGL’s rise in popularity has created a need for a new format that can represent rich data, and requires only minimal extra processing to be rendered [56].

glTF, the GL Transmission Format, is an upcoming runtime asset format specification, which aims to fill above need by directly loading data to WebGL buffers and designed to support rich data [56]. “Typed Arrays” is another technology that aims to solve the same problem without the complexity of glTF. It is part of next version of ECMAScript 6 specification and similar to how arrays work in C, it has an efficient way to support fast loading of binary data in WebGL [57]. We are analyzing both these upcoming specifications to support data exchange in our application.

6. ACKNOWLEDGEMENT

The authors would like to acknowledge Oak Ridge National Laboratory for providing the LandScan (2011)™ High Resolution USA Population Dataset, which is copyrighted by UT-Battelle, LLC, operator of Oak Ridge National Laboratory under Contract No. DE-AC05-00OR22725 with the United States Department of Energy. The United States Government has certain rights in this data set. Neither UT-Battelle, LLC nor The United States Department of Energy, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of the data set. The authors would like to acknowledge the support extended by Dr. Devin White and Dr. Cheng Liu.
7. REFERENCES


GLOSSARY:

- Cesium: a virtual globe and map engine based on WebGL [46].
- Cluster controller(s): controls provisioning and management of a cluster (or group of Nodes) [18, 19].
- gLTF: an upcoming runtime asset format specification, which aims to directly load data to WebGL buffers, avoiding unnecessary processing [56].
- ExtJS: one of the leading JavaScript application frameworks that is being used to create our interactive visualization web application using MVVW architecture [41, 54].
- LandScan: a high-resolution population distribution database, developed at ORNL [1, 2, 58].
- MAAS: a quick and easy-to-use tool for transforming machines into a cloud [17].
- Node: an individual machine. This is different from Node.js technology.
- Node.js: a server-side JavaScript environment for easily developing scalable network applications using asynchronous or non-blocking I/O model [26, 28].
- MongoDB: one of the popular NoSQL solutions [20, 21].
- OpenStack: “is the world leading open-source cloud platform and provides all the components needed to build and deploy an operational open-source cloud” with compute, storage and network components [16].
- OpenStreetMap (OSM): a public domain global street network [1, 2, 58].
- Open Layers 3: a cutting edge and high-performance mapping library, being developed by OSGeo to provide rich user experience including 3D using HTML5, WebGL, Cesium, CSS3 and Closure [41, 22].
- Postgres: one of the leading RDBMS [20, 21].
- Region controller: Easy and quick management of resources in the cloud via Web UI [18, 19].
- Three.js: provides various common utilities and helper functions to make use of WebGL simpler [37].
- Typed Arrays: ECMAScript 6 specification that aims to support fast loading of binary data in WebGL, without the complexity of gLTF [57].
• Ubuntu Server: is the operating system of the cloud [16].
• WebGL: a JavaScript API for rendering rich and interactive graphics including 3D in a browser without the need for plug-ins [33].
• Web Workers: a HTML5 specification that brings threading to JavaScript [47, 49].
• Web Sockets: a HTML5 specification that provides persistent duplex connection between the server and the client [53].
• Web Storage:
  o Session Storage: a HTML5 technology used to solve the storage problems associated with cookies [38, 39].
  o Local Storage: another HTML5 technology for storing data in key / value pairs – but, data persists in the client beyond a page refresh [38].
  o Indexed DB: an upcoming HTML5 API that provides a high-performance NoSQL-like data store for storing large amounts of structured data [40].
From Global to Local: Big Data and Model Development for
Spatially and Temporally Scalable Freight Transportation

Guoqiang Shen

Author Note
Correspondence concerning this article should be addressed to Guoqiang Shen, Regional and City Planning, The University of Oklahoma, 257 Gould Hall, Norman, OK 73019, USA
Contact: guoqiangs@ou.edu
Abstract

Using big-data mining and integration, geographical information systems (GIS), Google Earth™, and common transportation assignment methods (i.e., all-or-nothing, system optimal), this paper first reports a novel model that combines various big-data databases into a complete time-series port-level global maritime freight origin and destination database. This paper then performs freight flow assignments between foreign ports and U.S. ports and investigates total and commodity specific freight flows. The total flows from all ports of a country to or from all U.S. ports are considered as U.S. imports or exports, or international trade between this country and the U.S. Similar flow aggregations are done at the regional, country, port levels. The patterns of U.S. international trade are then explored and visualized by world region, country, port, in total or by selected commodities for the period of 1997-2007. The results show that U.S. imports outpaced exports in the period. Norman America (Canada and Mexico), Asia, Europe, and South America were among the top world regions exchanging commodities with the U.S. China, Japan, Canada, Mexico, Venezuela were the top countries shipping commodities to the U.S. At the port level, Long-Beach and Los Angeles ports were the most important, followed by New York City Port, Houston, New Orleans, San Francisco, etc. Top world ports, such as Hong Kong, Shanghai, Singapore, Pusan were the most interactive foreign ports with U.S. ports. Panama and Suez were important canals in ocean freight movement to and from U.S. Finally, China, a top trade partner with the U.S., exported and imported most of its commodities to the U.S. through its five east port clusters or three eastern megacities.

Keywords: Global trade pattern, Maritime freight flow, Data mining and integration, Geographical information system, Ports
Introduction

International trade is physically realized as goods movement through a global transportation network. Although multiple transportation modes are used for international trade, maritime transportation is the predominant mode. Over 90% of international trade is carried by maritime vessels and 70% of maritime cargo shipping is realized by containerized transportation (Kite-Powell, 2001; Maritime Transportation and Shipping Talk, 2008). Compared to other modes, such as air, rail, and highway, maritime shipping is considered as the most energy-conservation and low carbon-emission mode for freight (Frankel, 1989; Corbett, 2004, Corbett and Fischbeck, 1997; Winebrake et al, 2007).

Seaports are very important factors in maritime transportation since most vessels carrying freight move through ocean waterways connecting these seaports. Globally, there are over 4,000 such ports engaging in international maritime shipment (Landfall Navigation, 2005). In the U.S., there are over 300 maritime ports and quite a few of them engage in international trade (Maritime Administration, 2012). Most maritime ports receive imported goods and dispatch exported goods. Some sea ports, especially the large ones, also transship goods to other ports.

The global growth of international trade of goods and services was approximately twice the growth of the world gross domestic product (GDP) during 1995 to 2005 (Business & Economics Research Advisor, 2007; World Trade Organization, 2008). Global freight flows to and from the U.S. have significant impacts on the U.S. economy, business, quality of life, and national security. The value of the U.S. international trade of goods in 2005 (or 2012) were $3,992 (or $3,821) billion, of which exports accounted for $1,646 (or $1,546) billion and imports for $2,346 billion ($2,275), with a trade deficit of $700 (or $729) billion (Nanto et al., 2009; Bureau of the Census, 2012).

Of the goods imported into the U.S. in 2007, 52% (by value) were via waterborne transportation, 21% by air transportation, 15% by truck transportation, and 5% by rail transportation. Of the goods coming into the U.S. by water mode, 59% billion came into the country in containers. Containerized or non-containerized imports include both finished goods and intermediate parts or products. The increasing containerized trade is evidenced by the growing number of containers (of all sizes) entering the U.S. ports. In 2008, the port of Los Angeles accounted for 20.1% container volume (TEU), Long Beach for 16.4%, New York and New Jersey for 14.2%, Savannah, GA for 7.5%, and Norfolk, VA for 5.6%. The top three ports accounted for half of 2008 U.S. container volume. In 2007, 72% of U.S. import TEUs were from 10 countries with China (excluding Hong Kong) as the top trading partner, followed by Japan, Hong Kong, South Korea and Taiwan (Bureau of Transportation Statistics, 2009a, 2009b).
This research studies U.S. international maritime freight flows as proxies to the U.S. international trade patterns during 1997-2007. The paper starts with a concise survey on maritime freight transportation literature, followed by a review of major databases which contain relevant information of U.S. international freight. Then, it develops a data mining framework to integrate these databases to analyze the global freight flows to and from the U.S., hence the U.S. international trade patterns. Some important database mining, integration, and visualization issues, such as origin-destination (O-D) matrices, commodity codes, measurement units, spatial scales, and software integration in TransCAD™, ArcGIS™, and Google Earth™, are discussed. All-or-Nothing (AON) assignment using TransCAD™ was performed to generate and visualize the best global freight movements for U.S. international trade.

Literature review

International maritime goods are carried by shippers on vessels from origin ports in a country to destination ports in another country. These origin and destination ports are predominantly coastal sea ports or large inland water ports engaging in importing and exporting goods. These ports and the connecting waterways form the maritime network. Imported, exported, and transitional goods in the maritime network are also called maritime freight, often measured in value (e.g., $) or by weight (e.g., tons), is classified into different types of goods or commodities (e.g., Harmonized System). Ocean container is the dominant way for maritime freight, others include flatrack, platform, bulk, tank, etc.

The literature on international trade or its proxy measure of freight flows at port, local, national or regional levels is vast. Cullinane (2005), Robsion (2005), and Lowe (2005) performed extensive research on economic, environmental, safety, technical and policy issues related to intermodal freight. Limão and Venables (2001), Anderson and Wincoop (2004), Liu and Xin (2011) and Yip (2012) explored spatial, temporal, and transport cost effects on trade volumes, patterns for total and commodity-specific international trade (e.g., grain) using econometric modeling.

Maritime freight flows and transportation networks were studied by Hayuth and Fleming (1994) on seaports strategic location, centrality, and accessibility; by Montes et al. (2012) on the evolution of the containerized and general cargo maritime routes before the 2009-2011 period using graph theory; by Fremont (2007) and Ducruet, Rozenblat, and Zaidi (2010) on multi-level maritime hub-and-spoke strategies for the world’s largest container liner Maersk and the Atlantic case; and by Ducruet, Lee, and Ng (2010) for maritime network centrality and vulnerability for liner carriers in Northeast Asian ports for the period of 1996-2006. Maritime freight capacity and financing for ports and other infrastructure were examined by GAO (2002), TRB (2003), and Maloni and Jackson (2007). U.S. Maritime port selection and carrier choice for efficient intermodal freight were investigated by Klodzinski and Haitham (2003), Lou and Grigalunas (2003), and Melchow and Kanafani (2004).
Waters (1977), Notteboom and Rodrigue (2005), Ducruet (2007), Grobar (2008), Verhetsel and Sel (2009), and Talley, (2009) studied the regionalization, globalization, innovation, and importance of the global freight network and port cities for imports, exports, and transitions, and concluded that it is paramount important to develop world port cities and their associated hinterland so that they become the central nodes of global maritime freight flows. Deng, Lu, and Xiao (2013) however pointed out that demand, supply and value added activities in ports are important, but port supply and demand do not have significant positive effects on the ports’ regional economy.

Shashikumar and Schatz (2002), Lewis and Vellenga (2002), and Wang (2012) studied the regulatory aspect of ocean shipping, such as the impacts by the Ocean Shipping Reform Act (OSRA) of 1998 on the structure of the shipping market and the competition on major world trade routes by carriers. Heaver (1995), Cox (2000), and Wang, Ng, and Olivier (2004) explored governance, competition, and power dynamics at ports for freight supply and demand management and policy that influence port logistics capabilities and freight movements in the global maritime network.

Since 9/11, the concern of maritime security has been shifted to possible terrorist attacks to large freight ships and maritime logistics facilities and networks that could paralyze global maritime commerce (Frittelli, 2005, Nelson, 2012). Accordingly, the U.S. government and port authorities are taking steps to track the movement of maritime freight and hope to reduce potential risks (Voort et al. 2003; Chalk, 2008; Greenberg, 2009).

Despite the research on ocean freight networks, flows, regulations and associated environmental, economic, social, security, trade competition issues at port, local, national, and regional levels or for a specific liner or market, there is a lack of freight flow studies at the global level. This research is novel in that it explores the U.S. container and non-container imports and exports with all other countries in the world for the period of 1997-2007, visualizes the U.S. imports and exports as aggregated freight movement in the global waterway network in 2D and 3D, and highlights the aggregated ocean freight as a proxy to the U.S. international trade.

Maritime freight databases and issues

Many databases from government agencies and/or private organizations exist for U.S. global freight flows. In general, these databases can be classified into three groups: freight, sea port, and network databases. Freight in this study is primarily related to goods or commodities, including their classification, value, and tonnage. Sea ports are essentially the origins and destination infrastructure for international freight vessels to arrive and depart. Networks here refer to the inland and ocean routes that connect the sea ports for freight movement from one port to another.
Freight databases

They are mostly import and export freight flow data with attributes such as commodity codes, units, O-D matrices, and transportation modes, tons, and values. Primary freight databases containing valuable information of U.S. international freight flow are summarized below.

Freight Analysis Framework database version 3 (FAF3) is an open database, which is produced and updated by the Federal Highway Administration (2012) for U.S. international freight flow estimates for 2002 and 2007 and forecasts from 2010 to 2040 by five-year interval.

Maritime Administration Database (MAD) is developed by the U.S. Maritime Administration (2012). MAD provides the maritime freight flow information from 1997 to 2007 by “U.S. custom ports” and “trade partners” separately. All imports and exports in MARAD are presented by tonnage or container units. However, MAD does not classify flows into different commodities and have some port-to-port O-D information missing.

USA Trade Online (2012), an official source for U.S. merchandise trade data, is provided by the Foreign Trade Division of the U.S. Bureau of the Census. This database provides freight data from 1992 on a monthly frequency. It contains 6-digit North America Industry Classification System (NAICS) and 10-digit Harmonized Commodity Description and Coding System or Harmonized System (HS) by World Customs Organization (2012). As for O-D pairs, it collects the freight flow between foreign countries and U.S. ports.

Navigation Data Center (NDC) database by U.S. Army Corps of Engineers (2012a) supplies a more completed collection of maritime freight data, which comprises the information on foreign cargoes, facilities, and other correlative sources. The yearly maritime freight flow information for 1997-2007 was available for this research. Furthermore, the O-D pairs in NDC database are port-to-port pairs, which include over 200 U.S. ports and over 1,800 foreign ports. Lock Performance Monitoring System (LPMS) by U.S. Army Corps of Engineers (2012b) was used to classify commodities.

Port databases

Since the global freight flow is mostly shipped between maritime ports, it is important to study port-to-port freight movement as a way to understand international trade. Here, the geographic information of maritime ports is imperative. Two databases, NDC and World Port Index (WPI) by Landfall Navigation (2005), contain the geographic locations and other attributes for major U.S. ports and foreign ports.

Port locations in NDC (and its Schedule K files) are defined by longitudes and latitudes for most U.S. ports. However, NDC only provides very limited geographic information for foreign ports. Of the 1,813 foreign ports listed in the NDC, only a few of them have their longitudes and latitudes. Comparing with the NDC database, the WPI database has 4,043 port records with longitudes and latitudes. The 2005
version of WPI contains the geographic information for all U.S. and foreign ports. Since this research focuses on maritime transportation, thus a port in this paper refers to a maritime port.

Network databases

The NDC database contains some but very limited information of the global waterway, in which it only has very few ocean routes near U.S. coasts or inland waterways. By contrast, Oak Ridge National Laboratory (ORNL) (2009) provided a comparatively complete coverage of the global waterway network. In fact, the ORNL intermodal network unites the global waterways with U.S. domestic and intermodal networks and points.

The above maritime freight databases, together with a few other general international trade databases, are summarized into Table 1. The top 5 publically free databases are the most important, especially the top 4, and were used in this study. As the table shows, the databases have different origin-destination (O-D) resolutions (e.g., ports, cities, metropolitan areas), are classified in various commodity systems (e.g., HS, SCTG), are in diverse time horizons (e.g., month, yearly), and cover different time periods (e.g., 1960 to 2012, 2009). Some databases contain geographic and mode information, while other do not. In order to get an integrative target database with necessary attributes to study U.S. maritime freight flows, and hence, U.S. international trade patterns, a data mining framework is constructed and some important data integration and interaction issues are discussed.

Table 1.

<p>| Summary table of common freight databases for U.S. international trade |</p>
<table>
<thead>
<tr>
<th>Database Name</th>
<th>O-D Level</th>
<th>Commodity Code</th>
<th>History</th>
<th>Frequency</th>
<th>Unit</th>
<th>GIS Files</th>
<th>Cost</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORNL Transportation Networks</td>
<td>None</td>
<td>None</td>
<td>2009</td>
<td>None</td>
<td>None</td>
<td>Yes, Shape (point, line)</td>
<td>Free</td>
<td>Highway, Rail, Water, Air</td>
</tr>
<tr>
<td>Navigation Data Center (NDC) water networks</td>
<td>U.S. ports, foreign ports</td>
<td>None</td>
<td>1997 up to 2010</td>
<td>Yearly</td>
<td>None</td>
<td>Yes, Shape (point, line)</td>
<td>Free</td>
<td>Waterway</td>
</tr>
<tr>
<td>Maritime Administration</td>
<td>U.S. ports, foreign ports</td>
<td>6 digit, LPMS</td>
<td>1997 up to 2010</td>
<td>Yearly</td>
<td>Ton, Value</td>
<td>None</td>
<td>Free</td>
<td>Waterway</td>
</tr>
<tr>
<td>World Port Index (WPI)</td>
<td>World ports (lon and lat)</td>
<td>None</td>
<td>Up to 2012</td>
<td>None</td>
<td>None</td>
<td>Yes, Shape (point)</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>U.S. Trade Online</td>
<td>U.S. ports, foreign countries</td>
<td>10 digit, HS</td>
<td>1992-Now</td>
<td>Monthly, Yearly</td>
<td>Ton, Value, TEU</td>
<td>None</td>
<td>Fee</td>
<td>Waterway, Air</td>
</tr>
<tr>
<td>Global Insight - Transsearch</td>
<td>U.S. states, counties, BEA, foreign countries</td>
<td>2 digit, STCC</td>
<td>1995-2030</td>
<td>Monthly, Yearly</td>
<td>Ton, Value</td>
<td>Yes, Shape (point, line)</td>
<td>Fee</td>
<td>Highway, Rail, Water, Air</td>
</tr>
<tr>
<td>WISER Trade Foreign Database</td>
<td>U.S. states, foreign countries</td>
<td>6 digit, HS, SITC</td>
<td>1999 up to 2012</td>
<td>Monthly</td>
<td>Ton, Value</td>
<td>None</td>
<td>Fee</td>
<td>None</td>
</tr>
<tr>
<td>PIERS Global Intelligence Solutions</td>
<td>U.S. ports, cities, counties, metropolitan areas, states,</td>
<td>10 digit, HS</td>
<td>1992-Now</td>
<td>Daily</td>
<td>Ton, Value, TEU</td>
<td>None</td>
<td>Fee</td>
<td>Highway, Rail, Water, Air</td>
</tr>
<tr>
<td>Freight Analysis Framework (FAF3)</td>
<td>U.S. states, metropolitan areas, port districts, world regions</td>
<td>2 digit, SCTG</td>
<td>1998-2040</td>
<td>Yearly</td>
<td>Ton, Value</td>
<td>Yes, Shape (point)</td>
<td>Free</td>
<td>Highway, Rail, Water, Air</td>
</tr>
<tr>
<td>Commodity Flow Survey (CFS)</td>
<td>U.S. states, metropolitan areas</td>
<td>5 digit, STCC, SCTG</td>
<td>1983-Now</td>
<td>Every 5 years</td>
<td>Ton, Value</td>
<td>None</td>
<td>Fee</td>
<td>Highway, Rail, Water, Air</td>
</tr>
<tr>
<td>North American Transportation Atlas Data</td>
<td>U.S. state, Canada, Mexico</td>
<td>2 digit</td>
<td>1998 up to 2012</td>
<td>Yearly</td>
<td>Ton, Value</td>
<td>Yes, Shape (point, line)</td>
<td>Free</td>
<td>Highway, Rail, Water, Air</td>
</tr>
<tr>
<td>Foreign Trade Statistics</td>
<td>Port, states, district, foreign countries</td>
<td>6 digit, SITC</td>
<td>1960 up to 2012</td>
<td>Yearly</td>
<td>Ton, Value</td>
<td>None</td>
<td>Fee</td>
<td>Waterway, Air</td>
</tr>
<tr>
<td>U.S. International Trade Commission</td>
<td>U.S. states, regions, foreign countries</td>
<td>4-10 digit, HTS, SITC</td>
<td>1989 up to 2012</td>
<td>Monthly, Yearly</td>
<td>Ton, Value</td>
<td>None</td>
<td>Fee</td>
<td>None</td>
</tr>
<tr>
<td>TradeStats Express™</td>
<td>U.S. states, regions, foreign countries</td>
<td>3 digit, HS, SITC</td>
<td>1999 up to 2012</td>
<td>Yearly</td>
<td>Value</td>
<td>None</td>
<td>Fee</td>
<td>None</td>
</tr>
</tbody>
</table>
The database review reveals that different databases have different attributes and quite few of them use a variety of units or formats. Also, the utility of attributes from each database varies. These observations call for data mining and integration to derive the target database with desired attributes. To derive this targeted database, a three-step data mining model is developed. The model framework is shown in Figure 1. The integration of selected databases for the targeted maritime freight flow is shown in Figure 2.

The three-step data mining and integration model is a cohesive system, and it constructs the data mining and integration flow from the top down, with three main phases: data filtration, data integration, and data interaction. The data filtration step is to select a valid set of data sources from all available sources. The data integration step is to implement specific integration techniques to build the target database for a specific purpose. In the data interaction step, the target database is used to provide data analyses, such as scenarios and forecasts or modeling and reporting.

**Figure 1.** Data Mining and Integration Framework

**Database integration issues**

In the three-step data mining and integration model, data integration step is the key. Data integration is the process of combining useful attributes residing at different sources for a necessary
attribute set of the target database, and the primary action is to map the sources (Hand et al., 2001). In applying data integration for this research, the following important issues are identified and addressed.

![Maritime Freight Database Integration Diagram](image)

**Figure 2.** Maritime Freight Database Integration

**Various spatial scales**

The issue of various spatial scales is mainly from the definition of O-D pairs in the source databases. For example, one database has O-D data at the levels of region, country, or state, while another database provides freight information at the scales of metropolitan or port or port district. Thus, there is a need to bridge freight data at different spatial levels into a common spatial scale good for the target database. This issue can be addressed by aggregating or disaggregating O-D matrices of different geographic scales suitable for the target database.

**Commodity code mapping**

There are different commodity code systems adopted in different freight databases, hence, commodity code mapping is needed in data integration. The frequently-used code systems include Standard Classification of Transported Goods (SCTG) by Statistics Canada (2010), Standard International Trade Classification (SITC) by United Nations Statistics Division (2008), LPMS and HS. Through the technique of commodity code mapping, which essentially matches lower classification goods with more digits (e.g., 6, 8, 10 digits) to higher classification commodities with fewer digits (e.g., 2, 4 digits), we
could bridge different commodity code systems. Table 2 illustrates an example of code mapping for cereal grains commodity among SCTG, LPMS, and HS codes.

<table>
<thead>
<tr>
<th>Table 2 Code mapping example (SCTG-LPMS-HS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCTG</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>2100</td>
</tr>
<tr>
<td>2200</td>
</tr>
<tr>
<td>2901</td>
</tr>
<tr>
<td>2902</td>
</tr>
<tr>
<td>2903</td>
</tr>
<tr>
<td>2904</td>
</tr>
<tr>
<td>2909</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Various units

Three units are commonly used in maritime freight transportation, namely value ($), tonnage (TON), and number of containers (TEU). These different units can be mutually convertible in theory, but hard to be accurate in practice. Given various types of commodities and often mixed containers and non-containers, it is hard to convert TON to TEU to $ or vice versa, so in practice, different conversion ratios from surveys and case studies are often used.

Missing and mismatched data

The missing data and mismatched data are another two important issues for database integration (Brain, 2009; Shen and Aydin, 2012). For example, the NDC freight flow database, as one of the most complete databases, still contains many records with missing data on U.S. ports. Besides, the foreign port information in NDC and WPI databases to have lots of mismatched foreign port names, such as misspellings, multi-names, various language translation mistakes, or simply blanks.

Other data and software integration issues

In order to perform a spatial visualization and exploratory analysis in GIS, it is necessary to unify different coordinate systems and projections for various port (as node) and waterway (as link) databases. Also, data duplication of some records needs to be removed during integration. In addition, the vague definition of some attributes in databases also complicates data integration. Moreover, the selected geographic info for ports and links have to be seamlessly combined into one ocean network linking all U.S. ports to all world ports in TransCAD™5.0, ArcGIS™10.0, and Google Earth™ Pro, the GIS, transportation, and visualization software used in this research.

Highlights of U.S. international freight flows during 1997 to 2007

This research is not restricted to containerized maritime freight flow, but considers all kinds of international maritime freights in tonnage and value, including non-containerized commodities, hence, all
inbound and outbound maritime commodity flows, or simply maritime imports and exports. Figure 3 portrays the total U.S. maritime freight flows in imports, exports and transitions in tonnage respectively for the period of 1997 to 2007. It is clear that the U.S. trade deficit was widening almost consistently during this period primarily due to the faster increasing of inbound freight (imports) and flat outbound freight (exports) and relatively small and stable transshipment freight (transitions).

![Figure 3](image1.png)

**Figure 3.** U.S. maritime exports vs. imports in millions of tons 1997-2007

Figure 4 is a pyramid chart showing the U.S. imports (+) on the right-hand side and exports (-) on the left-hand side for the same eleven-year period. It not only echoes the overall U.S. import and export patterns in Figure 3, but also breaks down imports and exports by eight major world regions.

Several notable features can be seen in Figure 4. First, Asia was the only region having more exports than imports with the U.S. for most of the years and all other regions had more imports than exports. Second, Asia, North America, Europe, and South America had the most international trade with the U.S. while Australia & Oceania was the smaller yet more balanced trading region with the U.S. in the same period. Third, the U.S. imported much more from than exported to Africa, South America, and Central America & the Caribbean regions. Finally, the U.S. received a significant trade from the North American Free Trade Agreement (NAFTA) region. The Middle East & North Africa also shipped more goods to the U.S. than U.S. received from the region.

![Figure 4](image2.png)

**Figure 4.** U.S. maritime imports (+) and exports (-) from the world regions in millions of tons, 1997-2007
Figure 5 shows the total import and export flows between the U.S. and the eight world regions. While Asia, Europe, North America, and South America had similar total exports and exports with the U.S., their aggregated shipping distances and modes are quite different. For example, the freight between U.S. and NAFTA countries (mainly Canada and Mexico) was realized through shorter trans-border rail and highway networks, however, the freight between Asia and the U.S. was primarily realized via longer ocean routes. Aside from other supply chain constraints and comparative regional advantages in resources and commodities, the tradeoffs between longer distances but relatively cheaper labors in Asia with respect to closer proximity but higher labor cost in Canada and Mexico can be inferred here. Similar inferences can be made for international trades with South America and Africa. However, the similar total U.S. trade volumes with and aggregated distances from Asia and Europe implied that the comparative advantages of specializations and costs are all important.

Figure 5. U.S. total international trade by world regions, 2007

Figure 6 and Table 3 show the most important international trading countries in 2007. In Asia, China, Japan, South Korea ranked in the top three. Canada and Mexico were the top two in North America. In Europe, United Kingdom, Netherland, Spain, Italy, and Belgium were major traders. In South America, Venezuela, Columbia, Brazil were on the top. Finally, in Middle East, Saudi Arabia, Egypt, Iraq.
China and Japan (from Asia), Canada and Mexico (from North America), and Venezuela (from South America) are the large countries with large international trades with the U.S. during the period of 1997-2007. Table 3 shows the imports and exports between these countries and the U.S. in 2007. Canada led the way, followed by Mexico, Venezuela, China, and Japan. Interestingly, the U.S. had trade deficits (more imports than exports) with all these countries except Japan. Also, both Mexico and Venezuela
exported much more commodities classified as Misc. Edible Preparations (LPMS = 21) than their total imported goods from the U.S. LPMS=21 can be broken down 2101 = extracts etc. of coffee, tea or mate, roast chicory, 2102 = yeasts, dead sing-cell micro-org nesoai, baking powder, 2103 = sauces & preps, mixed condiments, mustard flour etc., 2104 = soups, broths & preps, homogenized comp food preps, 2105 = ice cream and other edible ice, with cocoa or not, and 2106 = food preparations nesoai.

Several additional freight flows can be observed from Table 3. First, a type of goods imported into the U.S. does not mean it cannot be exported from the U.S. For example, with China, the U.S. imported over 29 million tons of LPMS =70 and exported over 1.3 million tons of the same goods. The same can be observed with Japan (e.g., LPMS=51), with Canada (e.g., LPMS =22), with Mexico (e.g., LPMS = 23), and with Venezuela (LPMS = 24). The volumes traded can be lopsided like the case with China (e.g., imported 29,478,608 tons and exported 1,324,044 tons for LPMS = 70) or very similar in quantity like the case with Mexico (e.g., imported 1,440,221 tons and exported 1,448,596 tons for LMPS 23). Second, the same commodity can be imported from and exported to multiple countries with different or similar volumes. For example, China exported 1,273,416 tons LPMS = 68 to the U.S., while Japan imported 1,963,149 tons of the same goods from the U.S. in the same year, however, the U.S. imported 2,742,157 tons of LPMS =22 from Mexico and exported 21,952 tons of the same LPMS = 22 to Venezuela. Finally, each trading partner seemed to specialize on a handful of commodities as evidenced by the dominantly large volumes of these commodities in total traded flows. For example, about 59.1% of the U.S. imports from and 56.4% of the U.S. exports to China were on LPMS = 70 and 52 and LPMS = 65 and 42. Similarly, the dominant trade commodities for Canada were LPMS 21 and 43 for imports and LPMS 10 and 4 for exports. For Venezuela, the most important goods were LPMS = 21 for imports and LPMS = 32 and 62 for exports.

Figure 7. 1997, 2002, and 2007 total imports by weight at selected U.S. sea ports
Figure 7 shows U.S. import freight flows at selected major U.S. sea ports in 2007. Houston was the largest port for U.S. imports in 1997, 2002, and 2007 for total ocean cargo weight shipped including containerized and non-containerized goods. Some ports played an important role both for imports and exports, for example Houston remained the top 1 port for total imported weight during 1997-2007. Similarly, the Port of South Louisiana occupied the first place for total exported weight during 1997-2007. And some ports made different contributions to both imports and exports; for example Long Beach and Los Angeles were significant in exports and imports during 1997-2007.

Table 4 illustrates a snapshot of U.S. port ranks in 1997, 2002, and 2007. While some ports were consistently ranked in the top 10, such as CA1-2, LA1-5, TX1-5, NY, others were ups and downs in the ranking during 1997-2007. Such port ranking dynamics was affected by many factors such as macro economy and micro port and shipping operation.

Table 4

<table>
<thead>
<tr>
<th>Year</th>
<th>Label</th>
<th>Top1</th>
<th>Top2</th>
<th>Top3</th>
<th>Top4</th>
<th>Top5</th>
<th>Top6</th>
<th>Top7</th>
<th>Top8</th>
<th>Top9</th>
<th>Top10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>Import</td>
<td>TX1</td>
<td>TX2</td>
<td>NY</td>
<td>TX3</td>
<td>OT</td>
<td>PA</td>
<td>LA1</td>
<td>TX4</td>
<td>TX5</td>
<td>LA2</td>
</tr>
<tr>
<td></td>
<td>Export</td>
<td>LA2</td>
<td>TX1</td>
<td>LA3</td>
<td>VA1</td>
<td>CA1</td>
<td>VA2</td>
<td>OR</td>
<td>CA2</td>
<td>AL</td>
<td>FL</td>
</tr>
<tr>
<td>2002</td>
<td>Import</td>
<td>TX1</td>
<td>NY</td>
<td>OT</td>
<td>TX4</td>
<td>TX2</td>
<td>CA1</td>
<td>LA2</td>
<td>TX3</td>
<td>CA2</td>
<td>LA5</td>
</tr>
<tr>
<td></td>
<td>Export</td>
<td>LA2</td>
<td>TX1</td>
<td>CA1</td>
<td>VA1</td>
<td>WI</td>
<td>WA1</td>
<td>CA2</td>
<td>OR</td>
<td>LA3</td>
<td>NY</td>
</tr>
<tr>
<td>2007</td>
<td>Import</td>
<td>TX1</td>
<td>NY</td>
<td>OT</td>
<td>TX4</td>
<td>CA1</td>
<td>TX2</td>
<td>LA2</td>
<td>CA2</td>
<td>TX3</td>
<td>LA5</td>
</tr>
<tr>
<td></td>
<td>Export</td>
<td>LA2</td>
<td>TX1</td>
<td>CA1</td>
<td>VA1</td>
<td>CA2</td>
<td>NY</td>
<td>LA3</td>
<td>LA4</td>
<td>W1</td>
<td>OR</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Code</th>
<th>U.S. Port Name</th>
<th>Code</th>
<th>U.S. Port Name</th>
<th>Code</th>
<th>U.S. Port Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>TX1</td>
<td>Houston Ship Channel, TX</td>
<td>LA3</td>
<td>Port of New Orleans, LA</td>
<td>OR</td>
<td>Port of Portland, OR</td>
</tr>
<tr>
<td>TX2</td>
<td>Corpus Christi, TX</td>
<td>LA4</td>
<td>Port of Plaquemines, LA</td>
<td>AL</td>
<td>Mobile Harbor, AL</td>
</tr>
<tr>
<td>TX3</td>
<td>Texas City Channel, TX</td>
<td>LA5</td>
<td>Calcasieu River, LA</td>
<td>FL</td>
<td>Tampa Harbor, FL</td>
</tr>
<tr>
<td>TX4</td>
<td>Beaumont, TX</td>
<td>CA1</td>
<td>Long Beach Harbor, CA</td>
<td>WI</td>
<td>Duluth-Superior, MN/ WI</td>
</tr>
<tr>
<td>TX5</td>
<td>Port Arthur, TX</td>
<td>CA2</td>
<td>Los Angeles Harbor, CA</td>
<td>NY</td>
<td>Port of New York</td>
</tr>
<tr>
<td>LA1</td>
<td>Port of Baton Rouge, LA</td>
<td>VA1</td>
<td>Elizabeth River, VA</td>
<td>OT</td>
<td>LA Offshore Oil Port</td>
</tr>
<tr>
<td>LA2</td>
<td>Port of South Louisiana, LA</td>
<td>WA1</td>
<td>Tacoma Harbor, WA</td>
<td>PA</td>
<td>Philadelphia Harbor, PA</td>
</tr>
</tbody>
</table>

Table 5 lists the top 10 foreign ports for the U.S. imports and exports by weight respectively in 1997, 2002, and 2007. For imports from the U.S. at foreign ports, Ras Tanura, Saudi Arabia took the first place in 1997, but Cayo Arcos Terminal became the top port by replacing Ras Tanura in 2005. Tokyo, Japan had been the top port for exports for the past eleven years. Wei Hai, China made to the top 10 ports in 2002. Other well-known world ports, such as Pusan, Shanghai, Singapore, Hong Kong in Asia, Antwerp, Rotterdam in Europe, and some in South America and Middle East are also ranked in Table 5.

In addition to these highlights, one notable feature of Table 5 is that a top 10 import port may or may not be a top 10 export port in different years and vice versa. However, a top 10 import or export port more likely remains in the top 10 in different years. This may be partially due to port specialization in
import or export, but not both or this may be linked to the locations of the original supply and the final demand of traded goods. Nevertheless, this feature is worthy of further investigation.

Table 5

<table>
<thead>
<tr>
<th>Year</th>
<th>Im/Ex</th>
<th>Top1</th>
<th>Top2</th>
<th>Top3</th>
<th>Top4</th>
<th>Top5</th>
<th>Top6</th>
<th>Top7</th>
<th>Top8</th>
<th>Top9</th>
<th>Top10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>Import</td>
<td>51727</td>
<td>30771</td>
<td>20152</td>
<td>99930</td>
<td>20184</td>
<td>20154</td>
<td>30100</td>
<td>30753</td>
<td>30704</td>
<td>30763</td>
</tr>
<tr>
<td></td>
<td>Export</td>
<td>58886</td>
<td>42157</td>
<td>58309</td>
<td>58840</td>
<td>42305</td>
<td>58023</td>
<td>06645</td>
<td>58029</td>
<td>58201</td>
<td>20199</td>
</tr>
<tr>
<td>2002</td>
<td>Export</td>
<td>58886</td>
<td>58309</td>
<td>06645</td>
<td>57000</td>
<td>58023</td>
<td>58201</td>
<td>42157</td>
<td>58000</td>
<td>58023</td>
<td>20199</td>
</tr>
<tr>
<td></td>
<td>Import</td>
<td>20152</td>
<td>51727</td>
<td>30700</td>
<td>99900</td>
<td>99930</td>
<td>72105</td>
<td>57035</td>
<td>14065</td>
<td>75300</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>Import</td>
<td>20152</td>
<td>51727</td>
<td>30700</td>
<td>99900</td>
<td>99930</td>
<td>72105</td>
<td>57035</td>
<td>14065</td>
<td>75300</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Export</td>
<td>58886</td>
<td>58309</td>
<td>06645</td>
<td>57000</td>
<td>58023</td>
<td>58201</td>
<td>42157</td>
<td>58000</td>
<td>58023</td>
<td>20199</td>
</tr>
</tbody>
</table>

*Ports are listed by port code and grouped by country with port code in () after each port.

Table 6 lists U.S. imports, exports, and differences (import – export) by 2-digit LPMS code for 1997, 2002, and 2007. First of all, the total U.S. imports increased more than exports fluctuated, leading an increased trade deficit of 147 million tons of imported goods. This finding at the commodity level reinforces Figure 3. Second, some imported and exported goods were consistently dominant over the years, such as LPMS = 21, 23, 53, and 70, with each over 30 million tons as imports in 1997, 2002, and 2007 and LPMS = 10, 24, 32, 62, and 63, with each over 27 million tons as exports in the same years. Third, some goods contributed a lots to the U.S. trade deficit, such as LPMS = 21, 23, 52, 53, and 70, while other goods balanced the trade deficit, such as LPMS = 10, 62, 63, and 65. Finally, some imported and exported goods added to the U.S. international trade volume significantly, such as LPMS 10, 21, 23, 68, and 70, while others only marginally, such as LPMS = 31, 41, 45, 48, 61, 66, 67, and 99,
Table 6
U.S. imports and exports by commodity for 1997, 2002, and 2007 (million tons except LPMS = 45 in 10K tons)

<table>
<thead>
<tr>
<th>Commodity Categories</th>
<th>LPMS</th>
<th>Im97</th>
<th>Ex97</th>
<th>Im02</th>
<th>Ex02</th>
<th>Im07</th>
<th>Ex07</th>
<th>I-E97</th>
<th>I-E02</th>
<th>I-E07</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cereals</td>
<td>10</td>
<td>10.18</td>
<td>78.37</td>
<td>16.67</td>
<td>43.25</td>
<td>37.51</td>
<td>55.73</td>
<td>-68.19</td>
<td>-26.58</td>
<td>-18.22</td>
</tr>
<tr>
<td>Misc. Edible Preparations</td>
<td>21</td>
<td>425.8</td>
<td>3.63</td>
<td>479.2</td>
<td>1.21</td>
<td>521.53</td>
<td>0.08</td>
<td>422.13</td>
<td>478.01</td>
<td>521.45</td>
</tr>
<tr>
<td>Beverages, Spirits &amp; Vinegar</td>
<td>22</td>
<td>24.34</td>
<td>7.15</td>
<td>34.28</td>
<td>7.8</td>
<td>61.39</td>
<td>13.06</td>
<td>17.19</td>
<td>26.48</td>
<td>48.33</td>
</tr>
<tr>
<td>Residues From Food Industries, Animal Feed</td>
<td>23</td>
<td>55.89</td>
<td>15.1</td>
<td>63.95</td>
<td>19.29</td>
<td>66.7</td>
<td>33.33</td>
<td>40.89</td>
<td>44.67</td>
<td>33.37</td>
</tr>
<tr>
<td>Tobacco &amp; Manuf. Tobacco Substitutes</td>
<td>24</td>
<td>18.53</td>
<td>27.89</td>
<td>24.06</td>
<td>28.87</td>
<td>11.54</td>
<td>33.61</td>
<td>-9.36</td>
<td>-4.81</td>
<td>-22.07</td>
</tr>
<tr>
<td>Organic Chemicals</td>
<td>29</td>
<td>4.27</td>
<td>2.17</td>
<td>7.67</td>
<td>2.76</td>
<td>18.19</td>
<td>2.1</td>
<td>2.1</td>
<td>4.91</td>
<td>16.09</td>
</tr>
<tr>
<td>Fertilizers</td>
<td>31</td>
<td>4.26</td>
<td>15.6</td>
<td>6.13</td>
<td>13.25</td>
<td>9.95</td>
<td>10.99</td>
<td>-11.34</td>
<td>-7.11</td>
<td>-1.04</td>
</tr>
<tr>
<td>Tanning/Dyeing Extracts, Dyes, Pigments, Paints, Varnishes, Putty, Inks</td>
<td>32</td>
<td>19.94</td>
<td>34.27</td>
<td>33.44</td>
<td>41.72</td>
<td>36.62</td>
<td>49.18</td>
<td>-14.33</td>
<td>-8.28</td>
<td>-12.57</td>
</tr>
<tr>
<td>Articles of Leather, Saddlery &amp; Harness, Travel Goods, Handbags, Guts</td>
<td>42</td>
<td>1.1</td>
<td>10.53</td>
<td>1.24</td>
<td>12.67</td>
<td>2.05</td>
<td>17.63</td>
<td>-9.42</td>
<td>-11.43</td>
<td>-15.58</td>
</tr>
<tr>
<td>Furskins &amp; Artificial Fur, Manufactures</td>
<td>43</td>
<td>25.13</td>
<td>5.72</td>
<td>33.69</td>
<td>3.66</td>
<td>41.97</td>
<td>4.14</td>
<td>19.42</td>
<td>30.30</td>
<td>37.82</td>
</tr>
<tr>
<td>Wood &amp; Articles of Wood, Wood Charcoal</td>
<td>44</td>
<td>21.8</td>
<td>14.41</td>
<td>15.51</td>
<td>11.82</td>
<td>12.3</td>
<td>18.97</td>
<td>7.39</td>
<td>3.7</td>
<td>-6.67</td>
</tr>
<tr>
<td>Cork &amp; Articles of Cork</td>
<td>45</td>
<td>0.31</td>
<td>0.28</td>
<td>0.49</td>
<td>0.15</td>
<td>0.69</td>
<td>0.79</td>
<td>0.03</td>
<td>0.04</td>
<td>-0.1</td>
</tr>
<tr>
<td>Manu. Straw, Esparto, Other Plating Materials, Basketware, Wickerwork</td>
<td>46</td>
<td>19.6</td>
<td>3.39</td>
<td>15.51</td>
<td>2.24</td>
<td>18.33</td>
<td>2.73</td>
<td>16.21</td>
<td>13.26</td>
<td>15.6</td>
</tr>
<tr>
<td>Paper &amp; Board, Articles of Paper Pulp</td>
<td>48</td>
<td>0.9</td>
<td>0.03</td>
<td>1.88</td>
<td>0</td>
<td>3.15</td>
<td>0.13</td>
<td>0.87</td>
<td>1.87</td>
<td>3.02</td>
</tr>
<tr>
<td>Printed Books, Newspapers, Pictures, Manuscripts, Typescripts &amp; Plans</td>
<td>49</td>
<td>7.02</td>
<td>0.64</td>
<td>5.28</td>
<td>0.47</td>
<td>4.43</td>
<td>0.37</td>
<td>6.38</td>
<td>4.81</td>
<td>4.07</td>
</tr>
<tr>
<td>Wool, Fine or Coarse Animal Hair, Inc. Yarns &amp; Woven Fabrics Thereof</td>
<td>50</td>
<td>3.79</td>
<td>9.13</td>
<td>4.77</td>
<td>6</td>
<td>8.71</td>
<td>7.98</td>
<td>-5.33</td>
<td>-1.23</td>
<td>-1.28</td>
</tr>
<tr>
<td>Cotton, Inc. Yarns &amp; Woven Fabrics Thereof</td>
<td>52</td>
<td>20.5</td>
<td>2.08</td>
<td>31.03</td>
<td>1.2</td>
<td>28.83</td>
<td>1.57</td>
<td>18.42</td>
<td>29.83</td>
<td>28.26</td>
</tr>
<tr>
<td>Veg. Textile Fibers Nesi, Yarns &amp; Woven Etc.</td>
<td>53</td>
<td>30.96</td>
<td>1.7</td>
<td>30.81</td>
<td>1.03</td>
<td>32.1</td>
<td>2.73</td>
<td>29.26</td>
<td>29.78</td>
<td>29.37</td>
</tr>
<tr>
<td>Man-Made Filaments, Inc. Yarns &amp; Woven Etc.</td>
<td>54</td>
<td>6</td>
<td>1.78</td>
<td>14.63</td>
<td>5.96</td>
<td>15.97</td>
<td>9.51</td>
<td>4.22</td>
<td>8.67</td>
<td>6.46</td>
</tr>
<tr>
<td>Man-Made Staple Fibers, Inc. Yarns Etc.</td>
<td>55</td>
<td>1.59</td>
<td>1.16</td>
<td>2.69</td>
<td>0.32</td>
<td>2.85</td>
<td>0.35</td>
<td>0.43</td>
<td>2.36</td>
<td>2.3</td>
</tr>
<tr>
<td>Articles of Apparel &amp; Clothing Accessories-Knitted or Crocheted</td>
<td>61</td>
<td>1.12</td>
<td>0.83</td>
<td>1.44</td>
<td>1.16</td>
<td>1.81</td>
<td>1.18</td>
<td>0.29</td>
<td>0.28</td>
<td>0.62</td>
</tr>
<tr>
<td>Articles of Apparel &amp; Clothing Accessories-not Knitted or Crocheted</td>
<td>62</td>
<td>0.25</td>
<td>27.91</td>
<td>0.25</td>
<td>26.07</td>
<td>0.15</td>
<td>32.86</td>
<td>-27.66</td>
<td>-25.82</td>
<td>-32.72</td>
</tr>
<tr>
<td>Made-up Textile Articles Nesi, Needlecraft Sets, Worn Clothing, Rags</td>
<td>63</td>
<td>0.08</td>
<td>43.88</td>
<td>0.1</td>
<td>47.94</td>
<td>0.08</td>
<td>54.77</td>
<td>-43.8</td>
<td>-47.84</td>
<td>-54.7</td>
</tr>
<tr>
<td>Footwear, Gaiters, &amp; The Like</td>
<td>64</td>
<td>1.43</td>
<td>8.39</td>
<td>1.28</td>
<td>8.31</td>
<td>0.8</td>
<td>7.74</td>
<td>-6.96</td>
<td>-7.03</td>
<td>-6.94</td>
</tr>
<tr>
<td>Headgear &amp; Other Parts</td>
<td>65</td>
<td>0.46</td>
<td>27.06</td>
<td>0.64</td>
<td>33.91</td>
<td>0.28</td>
<td>33.61</td>
<td>-26.6</td>
<td>-33.27</td>
<td>-33.33</td>
</tr>
<tr>
<td>Umbrellas, Sun Umbrellas, Walking-Sticks, Whips, Riding-Crops &amp; Parts</td>
<td>66</td>
<td>2.72</td>
<td>4.57</td>
<td>3.37</td>
<td>3.93</td>
<td>4.86</td>
<td>3.88</td>
<td>-1.85</td>
<td>-0.56</td>
<td>0.99</td>
</tr>
<tr>
<td>Prepared Feathers, Human Hair &amp; Articles Thereof, Artificial Flowers</td>
<td>67</td>
<td>0.97</td>
<td>21.01</td>
<td>0.86</td>
<td>13.32</td>
<td>0.73</td>
<td>11.45</td>
<td>-20.04</td>
<td>-12.45</td>
<td>-10.72</td>
</tr>
<tr>
<td>Articles of Stone, Plaster, Cement, Asbestos, Mica or Similar Materials</td>
<td>68</td>
<td>21.61</td>
<td>15.25</td>
<td>24.21</td>
<td>15.7</td>
<td>28.12</td>
<td>18.17</td>
<td>6.37</td>
<td>8.52</td>
<td>9.95</td>
</tr>
<tr>
<td>Glass &amp; Glassware</td>
<td>70</td>
<td>32.79</td>
<td>13.6</td>
<td>54.86</td>
<td>12.42</td>
<td>74.29</td>
<td>20.28</td>
<td>19.19</td>
<td>42.45</td>
<td>54.01</td>
</tr>
<tr>
<td>Business Services, Health, Fin./Insur. Legal/Real Estate, Hotels, Repairs</td>
<td>99</td>
<td>0.52</td>
<td>0.42</td>
<td>6.25</td>
<td>3.98</td>
<td>5.86</td>
<td>3.68</td>
<td>0.1</td>
<td>2.27</td>
<td>2.18</td>
</tr>
</tbody>
</table>

Note: Im97, Im02, Im07 or Ex97, Ex02, Ex07 or I-E97, I-E02, I-E07 = imports or exports or differences in 1997, 2002, and 2007 respectively.

U.S. international maritime freight flow movement

The U.S. trade patterns can be spatially visualized by displaying import and export freight flows on waterways connecting U.S. ports and foreign ports. No capacity limitation was assumed for each waterway link and route, and the shortest path was considered as the best route between any two sea ports. Under this assignment, all freight flows between two ports would be assigned to the shortest path connecting the pair.

Figure 8 – Figure 9 provide 2D snapshots of port-to-port freight flows between the U.S. and foreign ports in 2002 at the world, country, and commodity levels. The figures clearly show that Panama Canal, connecting Atlantic and Pacific Oceans, is the most important canal for U.S. maritime shipments, especially those from and to the U.S. East Coast ports, to reach Latin America and Asia. Panama Canal will become more vital after its expansion scheduled to be completed in 2014 (Knight, 2008). To a lesser extent, the Suez Canal, north-south connecting the Mediterranean and the Red seas, also handles heavy freight flows between Europe and countries around the Indian and western Pacific oceans and strategically influences the U.S. maritime commerce (Padya, Herbert-Burns, and Kobayashi, 2011).

Specifically, Figure 8 shows all the imported and exported U.S. goods and their freight flows on the global ocean network. Figure 8 also contains a close-up view of waterway freight flows to and from...
the U.S. maritime ports. Figure 9 presents two extreme cases – most and least imported commodities in 2002. Clearly, the freight origins and destinations were at quite different foreign ports and connected by various shortest routes to U.S. ports in the global maritime network.

**Figure 8.** International and domestic views of total commodity and imports and exports, United States, 2002

**Figure 9.** U.S. global freight flows for most and least tonnage commodities in 2002

Figure 10 portrays the maritime freight flows between China and the U.S. in 2007. More specifically, the top portion of Figure 10 shows the total U.S. imports from China east coast’s five port clusters, which are Bohai Rim port cluster (e.g., Tianjin, Qingdao), Yangtze River delta port cluster (e.g., Shanghai, Ningbo), Pearl River delta port cluster (e.g., Hong Kong, Shenzhen), Southwest Coast port cluster (e.g., Zhanjiang, Fangcheng, and Haikou), and Southeast Coast port cluster (e.g., Fuzhou,
FROM GLOBAL TO LOCAL: BIG DATA AND MODEL DEVELOPMENT FOR SPATIALLY AND TEMPORALLY SCALABLE FREIGHT TRANSPORTATION

Quanzhou, Putian and Zhangzhou), The bottom portion marks the freight flows from Yangtze River delta port clusters to U.S. The first three port clusters, when considered together with their surrounding cities and towns, can also be grouped into three megacity regions in China – Bohai megacity region, Yangtze River delta megacity region, and Pearl River delta megacity region. In 2005, these megacities generated about 72.0% of total international trade and 80.0% of the total sea port throughput in China.

3D visualizations of global freight flows of U.S. international trade were implemented in Google Earth™ Pro and shown in Figure 11, whose left and middle images present the total Sino-U.S. imports and exports in 2007. The total U.S. imports and exports freight flows the world in 2007 was shown in the right image of Figure 11.

![3D visualizations of global freight flows](image)

**Figure 11.** Google Earth 3D visualization of U.S.-China trade (left, middle) and U.S.-world trade (right), 2007
This paper conducted a concise review of maritime freight transportation databases, developed a data mining and integration framework, and performed some exploratory analyses and visualizations of the U.S. global maritime import and export freight. The research provided an alternative way to understand and highlight U.S. international trade patterns. The databases were identified and the data mining and integration model was developed and implemented for the target database with desired attributes. Important data integration issues, such as spatial scales, measurement units and their conversions, missing data and date mismatches were discussed. Selected U.S. international trade patterns at the world, regional, country, and port levels were summarized. Sample best maritime freight movement routes for all commodities and specific commodities were visually mapped in 2D. Sample freight movements for trade between the U.S. and China and the world were shown in Google Earth™ Pro in 3D.

This research can be improved in many ways. First, the freight flows were analyzed yearly, but seasonal, monthly, or weekly trends can be shown or highlighted for a better understanding of U.S. maritime imports and exports dynamics. Second, the international freight flows are very complicated in the real world; some ports play more important roles in the trade than others, indicating a hierarchic system of ports over time. Third, the mapping of countries, ports, and the maritime network is mostly 2-dimensional, while 3-dimensional dynamic and real-time visualization is definitely better in displaying and analyzing freight flows. Fourth, since the U.S. maritime import and export flows are shipped to and produced at demand and supply points within the U.S., the linkage of international maritime port-to-port flows to their demand and supply points through U.S. intermodal networks, i.e., highway, rail, and water, certainly warrants further research. Finally, it would be interesting to see simulations of this model for various special policy and what-if concerns.

References


FROM GLOBAL TO LOCAL: BIG DATA AND MODEL DEVELOPMENT FOR SPATIALLY AND TEMPORALLY SCALABLE FREIGHT TRANSPORTATION


Big Brother is Watching You... To Predict Crashes

Paul St-Aubin (corresponding author)
PhD Candidate, Department of Civil, Geological and Mining Engineering
Polytechnique Montréal
C.P. 6079, succ. Centre-Ville
Montréal, Québec, Canada H3C 3A7
Email: paul.st-aubin@polymtl.ca

Nicolas Saunier (corresponding author)
Associate Professor, Department of Civil, Geological and Mining Engineering
Polytechnique Montréal
C.P. 6079, succ. Centre-Ville
Montréal, Québec, Canada H3C 3A7
Phone: (514) 340-4711 x. 4962
Email: nicolas.saunier@polymtl.ca

Luis Miranda-Moreno
Associate Professor, Department of Civil Engineering and Applied Mechanics
McGill University
Room 268, Macdonald Engineering Building, 817 Sherbrooke Street West
Montréal, Québec, Canada H3A 0C3
Phone: (514) 398-6589
Email: luis.miranda-moreno@mcgill.ca
Abstract

The age of Big Data is here and many industries have already started embracing it. The transportation industry stands much to gain from large-scale data analysis due to the complexity and pervasiveness of transportation in daily life, which promises smarter roads and a better understanding of our transportation needs and environment. But this inertia is also one of the greatest challenges to big data adoption initiatives. Transitionary technologies may, however, provide the answer to kick-start this migration today. This paper presents, in detail, a practical framework for implementation of an automated, high-resolution, video-based traffic-analysis system, particularly geared towards traffic flow modelling, behavioural studies, and road safety analysis. This system collects large amounts of microscopic traffic flow data from ordinary video cameras and provides the tools for studying basic traffic flow measures as well as more advanced, proactive safety measures. This paper demonstrates the process step-by-step illustrated with examples and applies it to a case study of weaving zones across a large set of data at a number of Québec roundabouts. In addition to providing a rich set of behavioural data, the analysis suggests a relationship between flow ratio at the weaving zone and safety, between lane arrangement and safety, and is inconclusive about the relationship between approach distance and safety.

Keywords: road safety, computer vision, collision prediction, motion pattern learning, interactions
1 Introduction

Affordable computing and flexible and inexpensive sensor technology are transforming the current practice and methods for traffic data collection, monitoring and analysis: big data is changing how we interact with our environment and approach problem solving tasks in the field of transportation. This should come to no surprise as the complexity of urban mobility complexity and the pervasiveness of geo-location devices in daily life lend themselves naturally to large data sets. In this context, the use of mobile and/or fixed video sensors for traffic monitoring and data collection is becoming a common practice not only for freeways but also for urban streets. This opens up possibilities for more dynamic traffic load balancing and congestion easing of road networks and in return provides researchers with participatory network usage data collection. This new situation in which traffic data is being collected intensively demands more intelligent and advanced methods for traffic data analysis; it is then not surprising that computer vision techniques have gained popularity given the potential of transforming the existing video-based traffic monitoring infrastructure into a highly detailed traffic data collection tool to identify and study traffic behaviours.

One such behavioural study application is in proactive road safety diagnosis. This has been a long-standing goal in the field of transportation safety. Traditional statistical methods applied to accident data require long observation periods (years of crash data): one must wait for (enough) accidents to occur. Beginning in the 1960s, attempts were made to predict collision rates based on observations without a collision rather than historical accident records [Perkins and Harris, 1968]: these methods are now termed surrogate safety methods. The traffic conflict technique was one of the earliest methods proposed which entailed the observation of quasi-collision events: situations in which road users were exposed to some recognizable risk (probability) of collision, e.g. a “near-miss”. However, several problems limited their adoption: the manual data collection method is costly and may not be reliable, and the definition and objective measurement of these events are difficult [Chin and Quek, 1997].

Today, with technological improvements in computing power, data storage, sensor technologies, and advances in artificial intelligence, these issues are quickly being addressed. This research presents the application of a video-based automated trajectory analysis solution which combines the latest advances in high-resolution traffic data acquisition [Saunier et al., 2010] and machine learning methods to model and
predict collision potential from relatively short, but extremely rich traffic data. This data is typically obtained from ordinary video data via computer vision from a camera situated at 10 m or more above the roadway [Jackson et al., 2013]. This trajectory data consists of position and velocity measurements of road users captured 15 to 30 times per second to a relatively high degree of accuracy. This amounts to several million individual instantaneous measurements over the period of one day at a typical site.

This high-resolution data permits the measurement of precisely defined instantaneous surrogate safety measures identifying collision probability. One such measure is time-to-collision (TTC) which measures the time remaining at any given instant to some collision point in the future defined by a collision course with another road user. This measure is useful as it provides the remaining time road users have to react to and avoid potential collisions. Higher TTCs are generally considered safer, though the precise link has yet to be validated. However, this measure relies on motion prediction hypotheses to identify collision courses. The traditional approach is to use constant velocity projection [Amundsen and Hydén, 1977] [Laureshyn et al., 2010] (situations in which road users fail to correct their course or even turn), which is the motion prediction method most frequently used, without any justification. This approach does not natively provide a collision course probability, and it will not be suitable in situations where observed trajectories do no include constant velocity displacements: for example, turning lanes in an intersection and movements in a roundabout.

More advanced collision course modelling efforts are being developed, including motion patterns which represent naturalistic (expected) driving behaviour learnt from the same data set. This procedure provides several potential collision points and their probability as a function of both the characteristics of the specific site and the road users’ behaviour. The motion patterns, or the distribution of trajectories at a site and their probabilities, may be described discretely over time and space [St-Aubin et al., 2014] or with prototype trajectories [Saunier et al., 2007]. The motion and collision predictions are computationally intensive as they explore, for each pair of road users, at each point in time, all future positions in time and space (typically subject to a time horizon). Furthermore, interaction complexity and exposure increase exponentially as the number of simultaneous road users in a scene increases. For example, over the course of one day, a typical intersection can experience between 100 thousands and 100 millions of these instantaneous interactions, depending
This paper presents a complete automated system for proactive road safety analysis that can deal with large amounts of video data. To the authors’ knowledge, the presented system is the most comprehensive to be applied to such big data collected in the field for a real world traffic engineering study. A large video dataset was collected at more than 20 roundabouts in Québec to study road user behaviour and their safety. Camera views record data at more than 40 roundabout weaving zones, an area within the roundabout delimited by an entry and the next following exit. Each camera records 12 to 16 h of video on a given work day, which constitutes a dataset of over 600 hours of video data. Applying the proposed method to this large dataset yields considerable amounts of indicators, from individual road user measurements, e.g. speed, to individual interaction measurements, e.g. TTC, to aggregated indicators per road user or interaction, to aggregated indicators per site over time and space.

Analyzing such big data is a challenge of a magnitude that has never been undertaken before in driver behaviour and road safety research. It holds the key to understanding the processes that lead road users to collide, and to design and validate safety indicators that do not require accidents to occur. The approach will be demonstrated on this video dataset to identify roundabout characteristics that influence road safety.

The paper is organized as follows: the next section presents the methodology, with practical examples drawn from the roundabout dataset, which is then applied to about half of the collected data and various system outputs are presented, before the conclusion and discussion of future work.

2 Methodology

2.1 Overview

Figure 1 outlines the general data collection and analysis framework. For a given research mandate, factors are selected for testing and a set of video data is collected at a sample of sites with adequate representation of these factors, while controlling for as many other factors as possible. With scene data and camera calibration parameters, feature tracking can be performed to extract trajectories [Saunier and Sayed, 2006]. The trajectories are raw spatial-temporal position data of moving objects within the scene. This positional data is processed to obtain derived measures such as
speed, heading and acceleration. Finally, scene information can be added to obtain higher-level data, such as movements referenced by lane, conflict measures, and other high-level interpretation behavioural measures (specific to the study). With a large amount of potential contributing factors (e.g. site characteristics), it may be beneficial to apply site clustering techniques before initiating behavioural measure correlation.

![Data flow diagram showing the overview of the system.](image)

**Figure 1:** Data flow diagram showing the overview of the system.

### 2.2 Video Data

Road user trajectories are extracted from video data using a feature-based tracking algorithm described in [Saunier and Sayed, 2006] and implemented in the open
source project Traffic Intelligence¹.

2.2.1 Trajectories: Positions in Space and Time (x,y,t)

Trajectories are a series of points in Cartesian space representing the position of (the centroid of) a moving object (road user) at time \( t \) on a planar surface. Height \( z \) is usually not considered. Points are evenly spaced in time with a consistent \( \Delta t \) equivalent to the inverse of the framerate of the video, i.e. a measurement is done for each frame. Typical framerates for video are between 15 to 30 frames per second, providing 15 to 30 observations per moving object per second. The object (road user) itself is represented by a group of characteristic features spread over the object and moving in unison.

Three potential sources of error exist: parallax, pixel resolution, and tracking:

- **Parallax error** is mitigated by maximizing the subtending angle between the camera and the height of tracked objects. In practical terms this requires a high view or ideally a bird’s eye view, tracking objects with a small height to base ratio. Passenger cars are generally more forgiving in this respect than trucks or pedestrians.

- **Pixel resolution** determines measurement precision. Objects further away from the camera experience lower tracking precision than objects near the camera. Error due to pixel resolution is mitigated by placing study areas nearer to the camera and using high-resolution cameras, although increases in resolution offer diminishing returns of tracking distance.

- Finally, **tracking errors** may occur due to scene visibility issues or limits with current computer vision techniques, in particular to handle data association (e.g. attach the trajectories to the right objects when they occlude each other). These erroneous observations have to be rejected or reviewed manually.

Depending on the steps taken to minimize tracking errors, feature-based tracking functions best over study areas of 50-100 m in length with high-to-medium speed, low-to-medium density flows. A sample of road user trajectories is presented as they are tracked in image space in Figure 2.

¹https://bitbucket.org/Nicolas/trafficintelligence/
2.2.2 Derived Data: Velocity & Acceleration

Velocity and acceleration measures are derived through differentiation from position and velocity over time respectively. These are 2-dimensional vectors with a magnitude (speed and acceleration) and a heading.

It should be noted however that each successive derivation increases pixel precession error for that measure. A velocity measure requires twice as many pixels as a position measurement. Similarly, an acceleration measurement requires three times as many pixels as a position measurement. This type of error can be compensated for with moving average smoothing over a short window (e.g. 5 frames). At this time, acceleration measurements are still too noisy to be useful for instantaneous observations. Higher camera resolutions should solve this problem in future applications.

2.2.3 Size of Data

Feature tracking provides a microscopic level of detail. Individual observations measured at a single site over the course of a normal day typically register in the tens of millions. The sample size (number) of individual tracking measurements (positions, velocities, etc.) per hour \( n \) can be estimated with the equation

\[
   n = f Q d
\]

where \( f \) is the number of frames per second of the video, \( Q \) is the average hourly flow-rate, and \( d \) is the average dwell time of each vehicle in the scene (excluding full stops). Dwell time is affected by the size of the analysis area in the scene and the average speed. As such, the size of the analysis area needs to be carefully selected.

2.3 Complementary Data

With the exception of speed and vehicle counts, vehicle trajectories offer little insight without context. Complementary data about the scene is collected in order to perform traffic studies and for higher-level interpretation. This data includes a wide variety of scene descriptors and design geometry attributes characterizing the factors under study. Finally, a traditional inventory of contextual factors that may be related to the behaviours under study needs to be constructed and associated with each site. These include number of lanes, lane width, horizontal and vertical
Figure 2: Vehicle #304 is shown approaching vehicle #303 which is engaging the roundabout in the wrong direction demonstrating a frequent violation leading to a traffic conflict.

signalization, pedestrian facilities, the built environment, and upstream/downstream distances to other intersections.

2.3.1 Analysis Area

The analysis area is a bounding polygon which confines analysis to a particular region of the scene. This serves to i) reject areas of the image with unsatisfactory feature tracking (particularly at the edges of the video), and ii) confine analysis to a particular region. For a cross-sectional or before-after study, analysis areas should conform to the same region of the roadway as much as possible. An example of the analysis area is demonstrated in Figure 3.

2.3.2 Alignments

Trajectory clustering is an important preliminary step in scene interpretation. Trajectory clustering is an abstract representation of movements along prototypical paths through a scene, called alignments. This is the foundation for relating spatial position with road geometry and, in particular, position of moving objects in relation to lanes and sidewalks. The alignment is represented as a simple series of points with a beginning and an end, typically in the same direction as the majority of flows along this path. This process introduces a new coordinate system which maps a position of a moving object in Cartesian space to a position in curvilinear space.
\[(x, y) \rightarrow (l, s, \gamma). \tag{2}\]

where a point located at \((x, y)\) in Cartesian space is snapped orthogonally to the nearest position on the nearest alignment \(l\), and is represented by the curvilinear distance \(s\) along this alignment from its beginning and the offset \(\gamma\), orthogonal to this alignment, measuring the distance between the original point and its position snapped to the alignment. A second pass may be performed over a window of time less than the time users take to perform real lane changes to correct any localized lane "jumping" errors which frequently appear near converging or diverging alignments. These coordinates are useful for studying following behaviour, lane changes, and lane deflection.

Many approaches exist to trajectory clustering: while some methods are supervised, many more are unsupervised (e.g. k-means [MacQueen, 1967]). Manual trajectory clustering is labour intensive and potentially a source of bias, but allows for tight control of scene description and analysis oversight. Unsupervised clustering is systematic but naive as this form of clustering can only make use of trajectory data to infer spatial relationship. Manual clustering along a series of splines, called alignments, is chosen for its simple implementation and tight control over interpretation. A hybrid approach, which automatically refines spatial positioning of the manually defined alignments through traditional unsupervised clustering approaches, is considered for future improvements.

### 2.3.3 Network Topology

Once trajectories are clustered, a network topology is constructed in order to be able to intelligently propagate future possible positions of moving objects through the network. In simple networks (i.e. two alignments), these movements are implicitly defined simply by observing lane change ratios, but in more complex networks, such as the network shown in Figure 3, movements may involve multiple lane changes and therefore may require a more general approach. A recursive tree model is employed.

Alignment extremities are linked to other nearby alignments, creating diverging or converging branches, as are momentarily adjacent alignments. Alternatively, alignments which run in parallel over a distance of more than 15 metres are instead grouped into corridors over which lane changes may occur freely. This creates a series of links and nodes with implicit direction which can be searched to determine all possible future positions of a moving object inside of this network. This serves to
reduce processing times of spatial relationship calculations between objects (triage) and provides more intelligent interpretation of spatial relationships.

Figure 3: The partial trajectories and scene of a multi-lane roundabout with a complex configuration of lanes (the south and east approaches are not visible). The alignments are in pink, while the connectors are in cyan. Some sample trajectories are highlighted in light grey.

2.4 Measurement Definitions

2.4.1 General versus Specific Analysis Measures

Some measures are generalizable for all traffic studies using alignments, while others are not. General traffic measures include speed profiles, counts, lane changes, origin-destination matrices, and basic spatial relationships including conflicts. Other measures may be specific to the study and generally require high-level interpretation (HLI). This interpretation makes use of study-specific geometric information to gen-
erate custom measures. As such, this section will not cover these custom measures, and will instead focus on generalizable measures. However, an application of HLI calculations will be briefly presented in section 3.3.

2.4.2 Interactions

An interaction quantifies the spatial relationship between moving objects in a scene, as is depicted in Figure 2. At the most fundamental level, an interaction is defined as a pair of moving objects simultaneously present in a scene over a common time interval (also referred to as a user pair). We further define an instantaneous observation (i.e. in a given video frame) within this time interval as an interaction instant.

This interaction definition is a generic precondition for any safety-related event of interest. In many scenes, it will include events of widely varying relationship to safety. For example, the significance of an interaction between two vehicles separated from each other physically (e.g. via a median or a large building) may not be comparable to an interaction between two vehicles merely separated by a painted line because the probability that one of the vehicles comes into contact with the other vehicle is reduced in the case of the median. This may cause issues when comparing different scenes if the analysis areas are not drawn consistently and may increase the computational burden of with collision prediction.

One solution is to filter user pairs based on physical access and proximity. A network topology coupled with a driving distance horizon is proposed. This is not a perfect solution, however, as physical access isn’t necessarily a binary option. In our median example, it is still physically possible, although less likely, for vehicles to cross-over into an opposing lane and cause a collision, although this is something that could be modelled.

2.4.3 Motion Prediction

Safety is evaluated from the observations of all vehicle interactions, by predicting future positions to determine if they are on a collision course and to characterize that collision course. The potential for collision of all interactions is measured by predicting future positions of vehicles at every instant in time and examining i) situations of particular probability of collision (i.e. threshold) or ii) evolution of the probability of collision over a time series. Several motion prediction methods are proposed for study [Mohamed and Saunier, 2013]:
• **Constant velocity** is the classic motion prediction model, wherein vehicles are projected along straight paths at a constant speed and heading using the velocity vector at that moment in time. This model is the simplest but also makes the most assumptions: only one movement is predicted at every instant, it does not depend on the context (road geometry or traffic), and the natural (non-reacting) motion of a moving object is a straight path (not always true). These assumptions may be adequate for specific applications of the methodology, e.g. highways [St-Aubin et al., 2013]. The current implementation is based off of [Laureshyn et al., 2010].

• **Normal adaptation** uses the initial velocity vector at the prediction moment to project trajectories, but modifies the velocity vector to account for normal variation. This model benefits from a wider range of possible outcome velocity vectors, but otherwise suffers the same problems and makes the same assumptions as constant velocity. The implementation of normal adaptation studied is based off of [Mohamed and Saunier, 2013], using a acceleration maximum $\alpha$ of

$$\alpha = \pm \frac{2}{f^2}$$  \hspace{1cm} (3)

and a maximum steering parameter $\sigma$ of

$$\sigma = \frac{0.2}{f}$$  \hspace{1cm} (4)

where $f$ is the number of frames per second of the video.

• **Motion patterns** are a family of models which use machine learning to calculate future position likelihoods from past behaviour [Saunier et al., 2007, Morris and Trivedi, 2008]. This type of model is the most promising as motion prediction is probabilistic in nature and inherently models naturalistic behaviour. However, motion patterns are complex to implement and expensive to process. The type of motion pattern being studied for implementation is a discretized motion pattern [St-Aubin et al., 2014].

As illustrated in Figure 4, motion prediction is performed for each user pair over each interaction instant $t_0$ for a number of time steps of size $\Delta t$ between $t_0$ and $t_0$ plus some chosen time horizon. Each motion prediction may generate for two road
users a series or a matrix of collision points with a sum of probabilities inferior or
equal to 1.

Figure 4: Collision prediction space in \((x, y, t)\) over \(\Delta t\) steps based on the conditions
at \(t = t_0\).

2.4.4 Time-to-collision

Time-to-collision (TTC) is one of the most popular surrogate safety measures. It is
a method of quantifying proximity to danger. Time-to-collision measures the time,
at a given instant \(t_0\), until two road users collide, if they collide, based on the motion
prediction model. In the simplest form, e.g. constant velocity, time-to-collision is
the ratio of differential velocity and differential position. A TTC value of 0 seconds
is, by definition, a collision. TTC is particularly useful as it has the same dimensions
as some important traffic accident factors such as user perception and reaction time
and breaking time. Larger values of observed TTC thus provide greater factors of
safety for these driving tasks.

Time-to-collision is measured instantaneously: a new value of TTC may be com-
puted for every instant. Thus, a pair of users may have a time series of TTC observ-
ations evolving over time. Some efforts have been made to study these evolutions
[Saunier and Mohamed, 2014]. Other approaches have focused on quantile or thresh-
hold observations (i.e. counting the number of interactions with minimum TTC below
a threshold as in classical traffic conflict techniques [Svensson and Hydén, 2006]), or even to examine instantaneous risk and significance of TTC [St-Aubin et al., 2013].

A sample pair of road user trajectories (#303 and #304, Figure 2) and spatial relationships simultaneously existing over a time interval lasting 64 instants or just over 4 seconds is presented in Figure 5. In this scenario, vehicle #304 is approaching at high velocity vehicle #303 which is engaged in an illegal U-turn (in a right-hand roundabout, users are supposed to travel counter-clockwise around the centre island at all times). The differential velocity $\Delta v$, relative distance $d$, and corresponding time $t$ is measured for every instant. In a matter of just under 4 seconds, the differential velocity changes from 9.63 to 2.26 m/s while the relative distance changes from 28.57 to 9.57 m. For every interaction instant of this user pair, motion prediction is used to calculate resulting TTC under each motion prediction method. These predicted collisions and associated TTC measures are presented in Figure 6. Motion pattern prediction generates many more possible collision points than constant velocity prediction, though each of these points has a lower associated probability. When several potential collision points are predicted, the expected TTC $ETTC_i$ at time $t_i$ is calculated as the probability-weighted TTC average

$$ETTC_i = \frac{\sum_{j=1}^{m} TTC_{ij} \cdot Prob(collision)_{ij}}{n},$$

of all possible collision points indexed $j = 1..m$ that could be reached with probability $Prob(collision)_{ij}$ [St-Aubin et al., 2014].

It is clear from both this figure and the trajectories themselves that constant velocity and normal adaptation motion predictions are inadequate for roundabout conflict analysis: the trajectories share the same destination yet they are on a collision course only for a brief period of time with these prediction methods.

### 2.4.5 Post-encroachment time

While prediction models and TTCs relate to the collision potential, other surrogate safety measures aim to measure collision proximity from crossing, but not necessarily colliding movements. Trajectory data is detailed enough to provide gap acceptance time (GT) and post-encroachment times (PET). These are measures that broadly characterize how aggressively and close in space and time merging and crossing tasks, respectively, are performed. As such, there is generally only one of these measures for the entire common time interval of a pair of road users. Gap acceptance time and PET fall under the category of high-level interpretation measures as the calculation
Figure 5: Vehicle #304 is shown approaching vehicle #303 which is engaging the roundabout in the wrong direction. Spatial relationship measures $\Delta V$, relative distance $d$, and time stamp $t$ are labelled along the time series every eight frames between the two trajectories. Light grey lines join the two trajectories at common time frames for visualization purposes.

of these measures cannot be generalized for all traffic studies, in part because the behaviour does not apply to all types of traffic interactions, and, in the case of gap acceptance time, because the measuring method may vary from one type of geometry to another.

For the crossing zone defined by the intersection of the two trajectories of a pair of road users, the post-encroachment time measures the time between complete departure of the first arriving vehicle, and first arrival of the next arriving vehicle. If $PET = 0$, a collision has happened. As such, higher PETs should demonstrate safer behaviour, although not necessarily linearly. An alternative to PET is predicted PET (pPET) which is measured from motion prediction instead of direct observation.
Figure 6: Time series of TTC observations for different motion prediction methods for the interaction between vehicles #303 and #304. Points correspond to TTC for a specific collision point and lines are weighted average observations per instant. The expected evolution of the time series occurs with a slope of one second to one second when the TTC observation at a given instant holds true, i.e. vehicles do not correct their collision course and a collision ensues. [Mohamed and Saunier, 2013].

Gap acceptance time similarly measures arrival and departure of a road user at a common crossing zone, but in this case, the crossing zone occurs in-line during a merging task, usually followed by following behaviour.

2.5 Indicator aggregation over time and space

Instantaneous surrogate safety indicators may be aggregated over time for each interaction (or user pair), over a given time interval for several road users and over space. Indicator distributions are generally shaped like Gamma distributions across the literature [Ismail et al., 2010, Autey et al., 2012, St-Aubin et al., 2013]. Quan-
tifying collision risk based on any of the surrogate safety indicators is the remaining puzzle piece. Using a TTC threshold has been the traditional approach in traffic conflict techniques [Svensson and Hydén, 2006], correlating a number of interactions with minimum TTC below a threshold with an expected number of collisions, though this constitutes a significant loss of information [Saunier and Mohamed, 2014] and this introduces assumptions in the model. One recent approach proposed a shifted gamma-generalized Pareto distribution model [Zheng et al., 2014].

Nevertheless, some qualitative analysis is possible in some circumstances, for example with a continuous mass shift of a probability distribution function as demonstrated in Figure 7. This approach has been tried in some early applications of the methodology, e.g. in [Ismail et al., 2010, Autey et al., 2012, St-Aubin et al., 2013]. Figure 8 demonstrates three different TTC distribution aggregation methods as used to represent nearly 3 million TTC observations over the course of one day at a single site: i) all instantaneous indicator values (subject to over-sampling of low severity values as well as over-sampling by slower road users and longer corridors), ii) minimum value of time series per user pair, or iii) 15th percentile value of time series per user pair. The 15th percentile is a practical solution to ignoring outliers that influence the maxima.

3 Experimental Results

3.1 Data Size

Video was collected at 20 roundabouts using two types of camera, a security camera VIVOTEK with a narrow lens filming at 15 frames per second at a resolution of 800*600 and a consumer camera GoPro 2 with a wide-angle lens filming at 30 frames per second at a resolution of 1280*960. The cameras are mounted on a specially constructed mobile video-data collection system built for temporary, high-angle video data collection, with tamper-proof, weather-proof, self-contained features presented in [Jackson et al., 2013].

In these 20 roundabouts, video was recorded for 40 merging zones of varying lane configuration, geometry, land use, and traffic volumes across the province of Québec. The merging zone of the roundabout is defined as the portion of the ring intersected by an approach and an exit. There is generally one merging zone between every pair of adjacent branches. Video data at each site was taken on one mild summer workday from 6 AM to 7 PM or 10 PM and captures both peak traffic
hours [St-Aubin et al., 2013]. This yields a total of 600 hours of video data, 50% of which has been fully processed at this time (see Table 1 for more information).

The software used is the open-source Traffic Intelligence project [Saunier et al., 2010, Jackson et al., 2013], itself based on the computer vision platform OpenCV [Brahmbhatt, 2013]. This software provides the basic feature tracking (the algorithm presented in [Saunier and Sayed, 2006]), trajectory management and coordinate projection functionality as well as a few usage-specific tools such as correction for lens distortion, trajectory clustering, and basic motion prediction functions. Some of the more advanced analysis tools and techniques presented in this paper are under development and will be made available as their functionality is completed and validated.
At 30 frames per second, a data collection at an intersection over a period of 12 hours (e.g. 7 AM to 7 PM), over a driving distance of 50 metres and at an average driving speed of 30 km/h, and with an average hourly volume of 500 veh/h yields approximately 90,000 instantaneous moving object measurements per hour. Additionally, each of these observations can have anywhere between 3 to 100 feature tracks associated with it. The recommended number of features to aim for is roughly 15-20 per object over time, depending on the typical duration of time spent by a road user in the field of view: this yields manageable data sizes (roughly 500 MB of storage per hour of video) while maintaining an adequate level of data richness and object representation. Video storage needs will vary greatly by camera choice, resolution, framerate, and video encoding settings.

Figure 9 shows hourly number of user pairs observed versus traffic volume. Trends are evident, but contributing factors are not clear (probably a mix of several...
Table 1: Data details

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Roundabouts</td>
<td>20</td>
</tr>
<tr>
<td>Analysis Areas</td>
<td>41</td>
</tr>
<tr>
<td>Hours of Video Data</td>
<td>610</td>
</tr>
<tr>
<td>Estimated Total Traffic Volume</td>
<td>120,000</td>
</tr>
<tr>
<td>Disk Space (Video + Data + Overhead)</td>
<td>1.9 TB</td>
</tr>
<tr>
<td>Veh-km Traveled</td>
<td>8400 veh-km</td>
</tr>
<tr>
<td>Processed to Date</td>
<td>≈ 50 %</td>
</tr>
</tbody>
</table>

Lane arrangement indicators. This will need further study. The number of user pairs per hour should be linearly correlated with the number of interaction instants as is demonstrated in Figure 10. If they are not, it is possible that analysis areas across sites are not comparable, particularly for time-series analysis and aggregated TTCs.

Most of the analysis is conducted on a pair of dedicated consumer-grade high-performance machines (INTEL Core i7 3770k processor with 16 to 32 GB of memory), with parallelisation of some tasks and work offloaded to a computing cluster when acceleration is necessary. Feature tracking performance depends on video resolution and special post-processing requirements such as stabilisation or lens correction (for distortion). A typical one hour $800 \times 600$ video is processed with current consumer-grade hardware in about an hour. A typical one hour $1280 \times 960$ video with correction for distortion can be processed in about two hours. Basic analysis on one of these trajectory sequences takes between 5 minutes and 30 minutes, depending on traffic in the scene, while surrogate safety analysis, particularly motion patterns, is very sensitive to the interaction complexity of the scene and can typically take anywhere between 1 to 48 hours to complete.

3.2 Sample Surrogate Safety Analysis

A sample surrogate safety analysis of three of the sites is demonstrated in Figure 11. This shows trajectory tracks projected in and with respect to the scene, mean speed and heading, and spatial distribution of motion-pattern-predicted collision points with instantaneous probability $> 10^{-5}$ and $TTC < 1.5$ s.

Figure 12 demonstrates a cross-sectional comparison of TTC distributions based on motion prediction at constant velocity for 20 merging zones for two contributing factors, each using all interaction instants. These distributions are aggregated
directly from all TTC observations, they are not means of the distributions at each site. Kolmogorov-Smirnov tests are performed between the distributions to quantify non-parametric dissimilitude. In the first diagram, a cross-sectional comparison is made for merging zones situated nearer or further than 300 metres upstream from another intersection. When this distance exceeds 300 metres, the distribution mass appears to shift left except for a sharp increase in small TTC below 0.5 seconds. It is so far unknown whether this small concentration of low-TTC conflicts offsets all other increases in TTC. This comparison remains therefore inconclusive. In the second diagram, a cross-sectional comparison is made between merging zones with high approach traffic volume ratios and low approach traffic volume ratios, where $R$ is the flow ratio between approach volumes and total volumes at the merging zone. In this comparison, a clear and consistent mass shift is observed, suggesting that high approach traffic volume ratios contribute to safer merging behaviour in a roundabout.
3.3 Sample High-Level Interpretation Analysis

Some high-level interpretation measures are also compiled using the data sample (for the same 20 merging zones). Figure 13 shows the mean speed profiles, with the interval at mean ± one standard deviation, through the roundabout merging zone of the same 20 samples as previously used for surrogate safety analysis. Speed profiles are mapped, not as a unit of distance, but rather as a unit of curvilinear location relative to the start and end of the merging zone. This is done to account for the large variability in diameter of roundabouts and in the angle between successive approaches across the sites. The position measurement re-sampling method as described in section 2.2.3 is used here to correct for oversampling bias introduced from varying speed between road users. Mean speeds are generally consistent with those in the literature, but variation does occur by relative location and movement type. In addition, cross-sectional analysis as in Figure 13 uncovers even larger variations in mean speed profiles (not shown).
Figure 11: Sample spatial data and analysis at 3 selected sites from top to bottom. From left to right, diagrams demonstrate trajectory tracks (positions) in analysis area and with descriptive alignments, mean speed and heading on a regular grid, and spatial distribution of motion-pattern-predicted collision points with instantaneous probability $> 10^{-5}$ and TTC < 1.5 s. All coordinates in metres, north pointing upwards.

Finally, Figure 14 shows distributions of accepted gap times of approaching vehicles and corresponding roundabout vehicles at the same sites. In a cross-sectional analysis, this quantifies vehicle insertion aggressiveness. Smaller accepted gaps might be explained by more impatient drivers, typically symptomatic of high volumes of continuous flow inside the roundabout and long wait times at the approach. Figure 15 demonstrates platoon sizes (uninterrupted passage of sequential vehicles). Users already inside the roundabout are generally more clustered than users enter-
Figure 12: TTC distributions based on motion prediction at constant velocity across 20 merging zones for two testable factors, using all interaction instants: a) for the upstream merging distance, results are difficult to interpret in terms of safety, and b) for the flow ratio of approach/total flows $R$, results suggest that merging zones where the approach accounts for the majority of flows are safer.

4 Conclusion

This paper demonstrates the theory and practical application of large-scale, automated, proactive road safety analysis using computer vision. The reader is led step-by-step through the challenges and process of collecting, processing, and analyzing video data with examples along the way. It demonstrates an early implementation in the form of a spatial and cross-sectional analysis using a large data set of roundabout video data to test several contributing factors. In addition to a rich set of behavioural data, the analysis suggests a relationship between flow ratio and safety, between lane arrangement and safety, and is inconclusive about the relationship between approach distance and safety.

Several technical challenges were outlined, notably tracking error, quantified probability of collision from TTCs, and aggregation and sampling considerations, as they still require particular attention. It is expected that these issues will be addressed as processing and analysis tools become more accessible, more collaborators contribute solutions to the open-source software, and as techniques applied to
transportation issues become more sophisticated.

The full results of the study over all 600 hours of video data will be the subject and focus of future papers. More advanced tracking, error detection, motion prediction models, and trajectory clustering will also be the subject of further research.

5 Acknowledgements

The authors would like to acknowledge the funding of the Québec road safety research program supported by the Fonds de recherche du Québec – Nature et technologies, the Ministère des Transports du Québec and the Fonds de recherche du Québec – Santé (proposal number 2012-SO-163493), as well as the varying municipalities for their logistical support during data collection. Shaun Burns, a Masters student at McGill University was also instrumental in the collection of video data.
Figure 14: Accepted gap time at merging instant of approaching vehicles.

References


Figure 15: Platoon size (uninterrupted sequential flow) comparison between roundabout lanes and approach lanes.

Transportation Research Record: Journal of the Transportation Research Board, 2198:52–64.


A Probabilistic Urban Link Travel Time Estimation Model Using Large-scale Taxi Trip Data

Xianyuan Zhan
Lyles School of Civil Engineering
Purdue University
550 Stadium Mall Drive
West Lafayette, IN 47907, USA
Email: zhanxianyuan@purdue.edu

Satish V. Ukkusuri, Ph.D.
Professor
Lyles School of Civil Engineering
Purdue University
550 Stadium Mall Drive
West Lafayette, IN 47907, USA
Email: sukkusur@purdue.edu

Word Count: 5215 words + 1 tables + 6 figures
ABSTRACT

Accurate estimation and prediction of urban link travel times are important for urban traffic operations and management. This paper develops a probabilistic mixture model to estimate urban link travel times using large-scale taxi trip data with partial information. Unlike typical GPS trace data, the taxi trip data used in this study provides limited trip level information, in which only contains the locations of origin and destination, travel times and distances, etc. The focus of this study is to develop a robust probabilistic link travel time estimation model and demonstrate the feasibility of estimating network conditions using large-scale data with partial information. In the model, the path taken by a taxi is considered as latent and modeled using a multinomial logit distribution. The observed path data given the possible paths set and the mean and variance of the link travel times can be thus characterized using a mixture distribution. A solution approach based on EM algorithm is proposed to solve the problem. The model is tested on estimating the mean and variance of the link travel times for 30min time periods using data and a test network extracted from Midtown Manhattan, New York City. More robust estimation results are obtained owing to the adoption of the probabilistic framework.

Keywords

Link travel time estimation;
Trip based data;
Large-scale data analysis;
GPS-enabled taxi;
Urban networks
INTRODUCTION

Accurate estimation and prediction of urban link travel times are essential for various applications in urban traffic operations and management. Traditional approaches for urban link travel time estimation and prediction have largely relied on data sources from fixed sensors, such as loop detectors (Coifman, 2002; Zhang and Rice, 2003; Oh et al., 2003; Wu et al., 2004), automated vehicle identification (AVI) sensors (Park and Rilett, 1998; Li and Rose, 2011, Sherali et al., 2006), and Remote Traffic Microwave Sensors (RTMS) (Yeon et al., 2008), etc. All of these approaches require installing corresponding fixed sensors for data collection. The huge amount of cost associated with installing and maintaining the physical facilities of sensors limits the use of the previous approach only to major road segments or small transportation networks.

With the rapid development of pervasive computing techniques, the large-scale geo-location data has emerged as a new alternative for urban link travel time estimation and predictions. For instance, the Global Positioning System (GPS) devices installed in dedicated fleets of vehicles or in users’ mobile phones can be viable sources of data for monitoring traffic conditions in large cities (Herrera, et al. 2010). Commercial companies such as Inrix (Inrix Inc) have already gaining profit by collecting and utilizing “large-scale” historical traffic data from GPS-enabled vehicles or mobile phones. As an important component of public transportation system in urban areas, taxicabs equipped with GPS devices has been increasingly considered as an ideal ubiquitous sensors to monitoring the traffic states in urban transportation networks. Currently, installing GPS devises in taxicabs has become a common practice by local agencies of many cities, which is mainly used to locate taxis and track lost packages, etc. However, even though huge amounts of GPS data are generated and collected every day in big cities, they are rarely used for traffic monitoring purposes. There are two nice features about the GPS data generated by taxis. Firstly, data related to taxi movements are abundant. In New York City (NCY), there are 13,000 yellow medallion taxis serving 240 million passengers per year and transporting 71% of all Manhattan residents’ trips (NYCTL, 2012). In Hong Kong, there are 15,000 taxis by the end of 2013 and transporting more than 1,000,000 passengers every day (Government of Hong Kong, 2013). The huge number of taxi trips generate vast amount of
data every day, providing valuable information about the real-time traffic conditions in urban
transportation network. Secondly, the unique mobility feature of the taxi data enables the all-
time monitoring of urban transportation networks with large coverage areas but without the
need of installing any fixed sensors.

Due to technological limitations and privacy concerns from taxi operation agencies,
detailed trajectory data from taxis are seldom available, while the trip based data from taxis are
shared by local agencies, e.g. New York City Taxi and Limousine Commission (NYCTLC).
Unlike typical GPS trajectory data, such large-scale taxi trip data provide limited information,
which only contain the origin and destination coordinates, travel time and distance of a trip.
However, the extensive amount of data records compensates for the incompleteness of the data
and makes the link travel time estimation possible. Currently, there are limited researches on
estimating urban link travel times using large-scale trip based data. Most of the research on
travel time estimation uses data from GPS-equipped vehicles or mobile phones rely on the GPS
trajectory information (Herrera, et al. 2010; Herring et al. 2010; Hunter et al. 2009; Zheng and
Zuylen 2012). These models require the complete trip trajectories to be known, which is
inapplicable when data provides only partial information. Zhan et al. (2013) proposed a short-
term urban link travel time estimation model using the large-scale taxi trip data. However, the
model only provides point estimates for the hourly average of link travel times, which is
incapable of incorporating the variability of link travel times. Impact from traffic signals and
rapid changes of traffic states in congestion can introduce large variation in urban link travel
times. Improper modeling of the variance in link travel times will result in biased and unreliable
estimates.

The focus of this study is to develop a robust probabilistic link travel time estimation model
that is applicable to large-scale trip based data, and also have the ability to capture the
variability in short-term link travel time estimation. The goal of this study is to demonstrate the
potential and practical value of estimating network wide link states using trip based taxi data
in urban transportation operation and management applications.

The paper is organized as follows: the next section presents the methodology of the
proposed model, which includes the description of model assumption, model development and
a proposed solution approach. The third section presents the numerical results and validation of the model. The last section concludes the paper and provides several future extensions of the model.

**METHODOLOGY**

This section presents the proposed probabilistic link travel time estimation framework. We will first introduce several modeling assumptions, followed by the definition of notations and the detailed description of the probabilistic link travel time estimation model.

**Model Assumptions**

To reduce the modeling complexity, we first impose following assumptions to the link travel time estimation problem:

1. Link travel times are modeled using normal distribution. Although more realistic distributions can be applied, using normal distribution to model the link travel time will significantly reduce the model complexity and also provide several nice analytical features.

2. The delay at intersections caused by traffic signals are combined into the link travel times. Since only limited path level information is available, it is insufficient to separate intersection delays from the total travel times, thus we focus on the estimation of the mean and variance of the link travel times over a short time period.

3. Travel times are independent among links. Although certain level of correlation may exist among nearby links, especially during congested hours, the proposed model assumes independence of travel times between different links. Such correlation could be partly captured in the estimation process using the observed path travel time data.

4. The travel time when a taxi driver traversing part of the link is proportional to the distance he/she traveled on the link. This assumption is used to obtain the travel time the taxi drivers spent on starting and ending links of the trip.

5. Driver’s route choice based on utility maximization. The assumption is that each driver minimizes both trip time and distance, so that the driver can make more trips and thus have higher revenue. Further, the path cost perceived during the route choice decision making
for each driver is assumed based on mean link travel times and distances.

3 Notations

- $x_l$ Link travel time of link $l$
- $\mu_l$ Mean link travel time of link $l$
- $\sigma_l^2$ Variance of link travel time of link $l$
- $d_k$ Total distance of path $k$
- $y^i$ Actual trip travel time of trip observation $i$
- $\hat{t}^i$ Predicted trip travel time of trip observations $i$
- $R^i$ Reasonable path set of trip observation $i$
- $d^i$ Actual trip distance of trip observation $i$
- $z^i_k$ Latent variable denoting the use of path $k$ in observation $i$, $z^i_k \in \{0,1\}$
- $\alpha$ Vector of distance proportion parameter $\alpha = (\alpha_1, \alpha_2)^T$
- $\beta$ Vector of parameters associated with the path cost computation, $\beta = (\beta_1, \beta_2)^T$
- $x$ Vector of link travel times in the network, total number of links=$m$
- $y$ Vector of trip travel times, total number of observations=$n$
- $z$ Vector of latent variable $z = (z^i_k), \forall k \in R^i, i = 1, \ldots, n.$
- $\mu$ Vector of mean link travel times $\mu = (\mu_1, \ldots, \mu_m)^T$
- $\Sigma$ Variance matrix of link travel times, $\Sigma = diag(\sigma_1^2, \ldots, \sigma_n^2)$
- $D$ Set of all trip distances

4 Probabilistic Link Travel Time Estimation Model

The model proposed in this work adopts a probabilistic framework. In this model, link travel
time $x_l$ of link $l$ is assumed to follow a normal distribution $N(\mu_l, \sigma_l^2)$, with mean $\mu_l$ and
variance $\sigma_l^2$. For simplicity, we express the mean of all link times to be $\mu = (\mu_1, \ldots, \mu_m)^T$
and $\Sigma = diag(\sigma_1^2, \ldots, \sigma_n^2)$. The path travel time is hence modeled as the summation of a set of
link travel times, and the probability of the actual trip travel time $y_i$ of observation $i$ given
the use of path $k$ is:
\[ \alpha_1 x_0 + \alpha_2 x_D + \sum_{l \in k} x_l \sim N \left( \alpha_1 \mu_0 + \alpha_2 \mu_D + \sum_{l \in k} \mu_l, \left( \alpha_1 \sigma_0 \right)^2 + \left( \alpha_2 \sigma_D \right)^2 + \sum_{l \in k} \sigma_l^2 \right) \] (1)

where \((\mu_0, \sigma_0^2), (\mu_D, \sigma_D^2)\) represent the mean and variance for the trip starting and ending links. As a taxi driver only experiences a part of the total link travel times when traversing on the starting/ending links of the trip (picking-up/dropping-off passengers in the middle of the street), follow the Assumption 4 in the previous section, we introduce \(\alpha_1, \alpha_2\) to be the distance proportions that the taxi traverses on the starting and ending links. Then the travel time of the taxi on the starting/ending links can be modeled as \(\alpha_1 x_0\) and \(\alpha_2 x_D\). For simplicity, denote:

\[
g^i_k(\mu) = \alpha_1 \mu_0 + \alpha_2 \mu_D + \sum_{l \in k} \mu_l, \quad h^i_k(\Sigma) = \left( \alpha_1 \sigma_0 \right)^2 + \left( \alpha_2 \sigma_D \right)^2 + \sum_{l \in k} \sigma_l^2 \] (2)

Thus \(g^i_k(x) \sim N \left( g^i_k(\mu), h^i_k(\Sigma) \right)\), and the probability of path travel time \(y^i\) using path \(k\) is

\[
P \left( y^i | k, x \right) = P \left( y^i | k, \mu, \Sigma \right) = N \left( y^i | g^i_k(\mu), h^i_k(\Sigma) \right) \] (3)

Since the detailed trajectory information is unknown in this trip based dataset, the actual path taken by a taxi driver needs to be inferred. Given the origin and destination of a taxi trip, the size possible path set of a trip is typically huge in a large network. To reduce the size of the problem and make the short-term travel time estimation problem tractable, we first obtain candidate paths for each trip record by constructing an initial path set using \(k\)-shortest path algorithm (Yen, 1971). Only the paths with distances that are not significantly deviate from the observed trip distances will then be included in a reasonable path set \(R^i\) for model estimation. The probability of taken a particular path \(k\) is evaluated using a route choice model based on utility maximization (Assumption 5), which is formulated as a multinomial logit distribution:

\[
p^i_k(\mu, \beta, D) = \frac{\exp \left[ -C^i_k(\mu, \beta, d^i_k) \right]}{\sum_{s \in R^i} \exp \left[ -C^i_s(\mu, \beta, d^i_s) \right]} \] (4)

where \(d^i\) is the actual trip distance; \(C^i_k(\mu, \beta, d^i)\) is the path cost function. Here we assume the perceived path cost is based on the trip distance and the mean link travel times \(\mu\), rather than the actual trip travel time. This will significant reduce the model complexity, and is also more reasonable, as drivers are not possible to know the actual trip travel times before they
make route choice decision, however, they may perceive the mean link travel times of the network based on their experiences. Following Zhan et al. (2013), we model the path cost function \( C_k^i(\mu, \beta, d_k) \) as a combination of trip travel time and trip distance, which is given as:

\[
C_k^i(\mu, \beta, d_k) = \beta_1 g_k^i(\mu) + \beta_2 d_k
\]  

where \( \beta_1, \beta_2 \) are cost coefficients for trip travel time and distance. According to Zhan et al. (2013), the estimated values for \( \beta_1, \beta_2 \) from the same large-scale taxi trip dataset are given as 0.275/min and 1.563/mile.

After developing the path travel time distribution of taking a particular path and the corresponding route choice probability, the path travel time of trip observation \( i \) can hence be modeled as following finite mixture distribution:

\[
P(y^i|\mu, \Sigma, D) = \sum_{k \in R^i} \pi_k^i(\mu, \beta, d_k)P(y^i|k, \mu, \Sigma)
\]  

Finally, the overall mixture model of all observations given the set of link travel time parameters \( \mu, \Sigma \) takes the form of:

\[
H(y|\mu, \Sigma, D) = \prod_{i=1}^{n} \sum_{k \in R^i} \pi_k^i(\mu, \beta, d_k)P(y^i|k, \mu, \Sigma)
\]  

The parameter to be estimated are \( \mu, \Sigma, \) and \( y, D \) are observed from the data. The plate notation of above model can be represented in Fig. 1. In the plate notation, the large rectangle plate represents the repetition for parameters \( L \) and data observations \( D \). The squares indicate fixed parameters and the circles indicate random variables. Filled-in shapes indicate variables with known values. The indication \( \{R_i\} \) indicate the set of reasonable path set of observation \( i \), and \( R^i_k \) represents a specific path in the reasonable path set.
Model Estimation Using EM Algorithm

The proposed problem is a large-scale problem involving huge amount of taxi trips, with each contains a certain amount of reasonable paths. An expectation-maximization (EM) algorithm (Dempster et al., 1977; Bishop, 2006) is developed to efficiently estimate the model parameters. EM algorithm is a powerful tool for finding maximum likelihood solution for models involving latent variables. The EM algorithm contains two repeated updating steps: the E (expectation) step and the M (maximization) step. The E-step takes expectation over the latent variable using the current parameter values to remove latent variables from the formulation, and M-step re-estimate the model parameter by maximizing the expected value of the complete-data log likelihood. The EM algorithm is generally preferred over directly maximizing the likelihood function when dealing with the models such as the incomplete-data log likelihood function defined in (7), as it can produce more robust estimates and avoid singularities of the likelihood function in which a mixture component collapses onto a particular data point (Bishop, 2006).

Detailed discussion about the EM algorithm and its convergence to local maxima of the likelihood function, please refer to Dempster et al. (1977) and Bishop (2006). The development and description of the proposed EM algorithm is presented as follows:

Step 1: Initialize model parameters: $\mu^{old}$ and $\Sigma^{old}$.

Step 2: E-step:

The route choice for each trip observation can be perceive as a latent variable in the model. Let $z_k^i$ be binary variable that takes values of 0 and 1, with 1 suggests the path being
utilized, then

\[ P(z_k^i = 1) = \pi_k^i(\mu, \beta, d_k) \]  

Thus

\[ P(y^i|z_k^i = 1) = P(y^i|k, \mu, \Sigma) \]  

\[ P(y^i|z) = \prod_{k \in R^i} P(y^i|k, \mu, \Sigma)^{z_k^i} \]  

Using Bayes Theorem

\[ P(z|y, \mu, \Sigma) \propto \prod_{i=1}^n \prod_{k \in R^i} [\pi_k^i(\mu, \beta, d_k)P(y^i|k, \mu, \Sigma)]^{z_k^i} \]  

From this posterior distribution, the expected value over \( z_k^i \) can be computed as

\[ \mathbb{E}(z_k^i) = \frac{\sum_{s \in R^i} \gamma(s^i) \ln \pi_s^i(\mu, \beta, d_k)P(y^i|k, \mu, \Sigma) + \ln P(y^i|k, \mu, \Sigma)}{\sum_{s \in R^i} \gamma(s^i)P(y^i|s, \mu, \Sigma)} = \gamma(z_k^i) \]  

In above equation, \( \gamma(z_k^i) \) is evaluated using the current parameter values \( \mu^{old} \) and \( \Sigma^{old} \).

Step 3: M-step:

The expected value of the complete-data log likelihood function is given as

\[ Q(\mu, \Sigma) = \mathbb{E}_z[\ln P(y, z|\mu, \Sigma)] = \sum_z P(z|y, \mu, \Sigma) \ln P(y, z|\mu, \Sigma) \]

\[ = \sum_{i=1}^n \sum_{k \in R^i} \gamma(z_k^i) \ln \pi_k^i(\mu, \beta, d_k) + \ln P(y^i|k, \mu, \Sigma) \]

\[ = \sum_{i=1}^n \sum_{k \in R^i} \gamma(z_k^i) \left\{ -\beta_1 g_k^i(\mu) - \beta_2 d_k - \ln \sum_{s \in R^i} \exp[-\beta_1 g_s^i(\mu) - \beta_2 d_s] \right\} + \frac{1}{2} \ln (2\pi) + \frac{1}{2} \ln h_k^i(\Sigma) - \frac{[y^i - g_k^i(\mu)]^2}{2h_k^i(\Sigma)} \]  

The updated parameter is obtained by maximizing \( Q(\mu, \Sigma) \), it can be seen as solving
following constrained optimization problem:

\[
\max_{\mu, \tau} Q(\mu, \Sigma)
\]

s.t. \( \mu_i \geq t_{min} > 0 \)  

Note \( Q(\mu, \tau) \) is twice continuous differentiable, thus a wide range of constrained optimization algorithm can be used to efficiently solve above problem.

Step 4: Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

\[
(\mu^{\text{new}}, \Sigma^{\text{new}}) = \arg \max_{\mu, \Sigma} Q(\mu, \Sigma)
\]

\[
\mu^{\text{old}} \leftarrow \mu^{\text{new}}, \quad \Sigma^{\text{old}} \leftarrow \Sigma^{\text{new}}
\]

and repeat E-step and M-step in Step 2 and 3.

NUMERICAL EXPERIMENTS

Test Data and Network

The trip based data used in this research was collected by NYCTLC. The data contains the information of origin and destination geographical location, trip distance, trip duration and other related information. Around 30,000 to 50,000 daily trips are recorded in the entire year of 2013. In this study, we extract a week’s data (from 2013/10/7 to 2013/10/13) to test the proposed model. A 1175m × 1780m rectangle area in Midtown Manhattan is selected to serve the study region. The corresponding transportation network inside the study region is illustrated in Fig. 2, which contains 136 nodes and 254 directed links. This network includes highly congested road segments in Midtown Manhattan, such as 5th Avenue, 7 Avenue, Broadway. Severe congestions are expected to be observed in the estimation results. All taxi trip data with both origin and destination fall within the study region are extracted to test the proposed model.

To perform the travel time estimation, we split the extracted data into 30min time intervals, and the link travel times are estimated using all the data from a time interval. The choice of using 30min time interval is to guarantee enough data are available to perform travel time estimation, while keeping the length of time interval as short as possible. The amount of data
observed for each of the 30 min time interval over the study week is illustrated in Fig. 3. From the data, we observe as many as 1400 trip observations for weekdays and 800 observations for weekends within a time interval. Specifically, we select four time periods (9:00-9:30, 13:00-13:30, 19:00-19:30 and 21:00-21:30) to test the proposed model. The time period from 9:00 to 9:30 contains the maximum number of trip observations in weekdays, which represent the morning peaks. The 13:00-13:30 and 19:00-19:30 time periods correspond to another two smaller peaks in Fig. 3 and the time period of 21:00-21:30 is selected to test for off-peak situations.

Fig. 2 Test network and study region

Fig. 3 Number of observed trips for each 30 min time interval in the study region
Numerical Results

The proposed model is implemented in MATLAB and parts of the codes are compiled into C to improve the computation efficiency. Before running the numerical scenarios, the $k$-shortest paths ($k = 20$) for each nodal pair in the network are computed. This step is necessary, since it can avoid the expensive $k$-shortest path computation during estimation process. The reasonable path set of the data can be then efficiently obtained by directly utilizing the already computed $k$-shortest path set of the network. A threshold ratio $r = 25\%$ is also introduced to filter out unqualified data which path distance is deviated more than $1 \pm r$ compared with the observed path distance. The proposed EM algorithm has shown good convergence property in the numerical tests. The convergence plot for the expected value of the complete-data log likelihood $Q(\mu, \Sigma)$ and the incomplete-data log likelihood $LL(H(y|\mu, \Sigma, D))$ is illustrated using the example of Monday 9:00-9:30 scenario. Rapid convergence is achieved during the first few iterations, and it is observed that 50 iterations are sufficient to obtain convergent solutions for all the test scenarios. The entire estimation process can be finished within 15min using a 2.4GHz CPU laptop. The computation time can be further reduced by implementing parallel computing technique or using a more powerful computer.

![Convergence plot of the proposed EM algorithm for scenario: Monday 9:00-9:30](image_url)
The proposed model is tested on the four time periods for each day of the selected week. The validation results for all tests are presented in Table 1 in Appendix, which will be discussed in the following section. Due to the space limit, we only present the estimation result for a representative weekday (Wednesday) and a weekend (Saturday) in Fig. 5 and Fig. 6. The estimated mean link speeds (Fig. 5(a), Fig. 6(a)) instead of estimated means of link travel times are used to give a more intuitive representation of the results. Note that since the signal delay...
are included into the link travel time estimation, thus the value of presented mean link speed will be lower than the usual traversing speed people actually experienced while driving. The model estimation results are consistent with the highly congested expectation of the network in the study region. For weekdays, it can be observed that the entire test network is severely congested in 9:00-9:30 and 13:00-13:30 time periods. Most of the links in the network have low mean travel speeds. The results show that almost 60% of the links in the test network have mean speeds ranging from 2-10 mile/hour in both of the two time periods. The congestion has seen a trend of alleviation in 19:00-19:30 time period, as more links are observed to have higher mean speeds. During 21:00-21:30 off-peak time periods, the traffic condition is observed to be greatly improved, that almost 50% of links have mean speed ranging from 5-15 mile/hour. The situation for weekend is quite different from weekdays, where it is less congested during 9:00-9:30 time period, and then becomes congested in 13:00-13:30 and 19:00-19:30 time period. The traffic condition in these two time periods are still better than congested hours in weekdays, since there are about 60% of links have mean speeds ranging from 3-15 mile/hour.

We also present the normalized standard deviation (estimated travel time standard deviation divided by link length) as a measure of the uncertainty and the variability of the estimated link travel times, presented in Fig. 5(b) and Fig. 6(b). The reason for normalizing the estimated standard deviation of link travel times is to ensure it is comparable across different links. From statistical estimation perspective, the estimated normalized standard deviation is found larger for cases with few observations (e.g. 9:00-9:30 time period for weekend), as too few information is available to infer the model parameters. For time periods with higher number of observations, there is no obvious pattern for the normalized standard deviation of link travel times. This reflect the facts that the variability of link travel times are largely dependent on corresponding traffic condition during specific period. By estimating the variance of the link travel times, we are able to capture the variability of short-term link travel times, and have more robust interpretation of urban traffic network conditions.

Validation

Because the ground truth data is not available in this research, we validate the result by evaluating the model predicted path travel times against observed path travel times. Let
(\(\mu^*, \Sigma^*\)) be the estimated means and variances for the link travel times. The predicted path travel time for observation \(i\) is thus estimated as the path travel time with using the most likely path. Let \(K = \max_k \{\pi_k^i(\mu^*, \beta, d_k), k = 1, 2, ..., |R^i|\}\), then the predicted path travel time is given as:

\[
t^i = g_k^i(x^*) \sim N\left(g_k^i(\mu^*), h_k^i(\Sigma^*)\right)
\]  

(15) with the mean and variances estimated as

\[
E(t^i) = g_k^i(\mu^*), \quad Var(t^i) = h_k^i(\Sigma^*)
\]  

(16)

Different validation criteria are used to examine the quality of the proposed model, the validation results of all tested scenarios are presented in Table 1. To examine the overall fitting to the data, we present the log-likelihood of the model (\(\log H(y|\mu, \Sigma, D)\), denoted as \(LL\) in Table 1 in Appendix). In addition, we also evaluate absolute mean percentage error (\(MAPE\)) of the observed path travel times against the predicted mean values:

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{t^i - y^i}{y^i} \right| \times 100\%
\]  

(17)

Furthermore, since the predicted path travel time is estimated as a distribution (15), we also present the percentage of the observed path travel times that fall within the 95% and 99% confidence interval of the predicted path travel time distribution, denoted as \(P_{95\%}\) and \(P_{99\%}\). The criteria can be an important measure to examine the ability of the proposed model to explain the observed path travel times.

Reasonably good validation results are obtained from the test scenarios. From Table 1, it is shown that except for the time period of Monday 9:00-9:30, all \(MAPE\) values are below 30%. For some less congested time periods (fewer observations), the computed \(MAPE\) can be as low as 20%. Larger \(MAPE\) values are observed for congested hours (e.g. 9:00-9:30, 13:00-13:30), which is mainly due to the rapid change in traffic state during congestion within the 30min time interval. Another potential source of variation in travel times originated from the signal delay. In an urban transportation networks, intersection delay at a traffic signal sometimes can be even greater than the traversing time of the link itself, which introduces considerable level of uncertainty in the link travel time estimation. To address the uncertainty
generated from aforementioned sources and measure the explanatory power of the proposed model, we examine the $P_{95\%}$ and $P_{99\%}$ for the tested scenarios. The results show that for most of the tested scenarios, about 70\%~85\% and 75\%~92\% of the observed path travel times are fall within the 95\% and 99\% confidence intervals respectively given the predicted mean and variance of the path travel time. The high proportion of $P_{95\%}$ and $P_{99\%}$ suggests good explanatory power of the model.

CONCLUSION AND FURTHER EXTENSIONS

This study develops a probabilistic mixture model to estimate urban link travel times from the large-scale taxi trip data. The model only needs partial information provided in the data, in this case, the origin and destination location, trip travel time and distance. The path taken by the taxi is considered as latent and modeled using a multinomial logit distribution. The likelihood of the observed path data given the reasonable path set and the mean and variance of the link travel times can be then characterized using a mixture distribution. A solution approach based on EM algorithm is proposed to efficiently solve the problem. More robust estimation results are obtained owing to the adoption of the probabilistic framework.

Currently, there is no ground truth data available for validation, the model is validate through examine the goodness-of-fit to the observed data. Future research can be done to further verify the proposed model by either compare the estimates against speed data from loop detectors, or detailed trajectory information of taxi trips, which is collected by NYCTLC, but currently not available to researchers. It should also be noted that the proposed model is also applicable to trajectory data, since the intermediate trajectory points can be treated as the origin and destination pairs in the model. Using such more detailed data, the accuracy of the link travel time estimates can be greatly improved.

The proposed model has provided a very flexible probabilistic framework and several extension can be made to further improve the accuracy of the estimation. Using a more realistic link travel time distribution will greatly relax the restrictive assumption of the normal distributed link travel times. Recent literature has shown that the link travel times are more likely to follow a bimodal distribution due to the involvement of signal delay. Incorporating such class of realistic distribution can be helpful to better account for impact from signal delays.
Furthermore, the proposed model can be easily extended as a Bayesian mixture model by incorporating prior distributions $p(\mu, \Sigma)$ over the model parameters $\mu, \Sigma$. The Bayesian extension of the proposed model will provide more robust parameter estimates and more importantly, allowing for incorporating the historical knowledge about network link travel times using the prior distributions. The EM algorithm developed in this paper can be slightly modified to find the MAP (maximum posterior) solutions of the Bayesian version of the mixture model by simply changing the expected value of the complete-data log likelihood function $Q(\mu, \Sigma)$ in the M-step to $Q(\mu, \Sigma) + \ln p(\mu, \Sigma)$. By maximizing the modified objective function in M-step, the solution of the Bayesian estimates can be found. These further extensions of the proposed model would lead to a more accurate and robust link travel time estimation for urban traffic operation and management and fully utilize the abundant large-scale path-based taxi data available in big cities.

REFERENCES


# Appendix

## Table 1  Validation results for the tested scenarios

<table>
<thead>
<tr>
<th>Day</th>
<th>Criteria</th>
<th>9:00-9:30</th>
<th>13:00-13:30</th>
<th>19:00-19:30</th>
<th>21:00-21:30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>(N_{obs})</td>
<td>740</td>
<td>544</td>
<td>664</td>
<td>492</td>
</tr>
<tr>
<td></td>
<td>(LL)</td>
<td>-1817.67</td>
<td>-1375.00</td>
<td>-1247.96</td>
<td>-730.32</td>
</tr>
<tr>
<td></td>
<td>(MAPE)</td>
<td>31.39%</td>
<td>29.41%</td>
<td>23.95%</td>
<td>23.03%</td>
</tr>
<tr>
<td></td>
<td>(P_{95})</td>
<td>87.16%</td>
<td>84.38%</td>
<td>74.25%</td>
<td>69.31%</td>
</tr>
<tr>
<td></td>
<td>(P_{99})</td>
<td>90.94%</td>
<td>90.07%</td>
<td>81.63%</td>
<td>74.59%</td>
</tr>
<tr>
<td>Tuesday</td>
<td>(N_{obs})</td>
<td>776</td>
<td>526</td>
<td>510</td>
<td>563</td>
</tr>
<tr>
<td></td>
<td>(LL)</td>
<td>-1843.15</td>
<td>-1264.91</td>
<td>-1211.96</td>
<td>-935.85</td>
</tr>
<tr>
<td></td>
<td>(MAPE)</td>
<td>26.34%</td>
<td>25.91%</td>
<td>29.17%</td>
<td>20.52%</td>
</tr>
<tr>
<td></td>
<td>(P_{95})</td>
<td>81.70%</td>
<td>84.03%</td>
<td>81.18%</td>
<td>74.25%</td>
</tr>
<tr>
<td></td>
<td>(P_{99})</td>
<td>87.37%</td>
<td>88.78%</td>
<td>87.06%</td>
<td>82.42%</td>
</tr>
<tr>
<td>Wednesday</td>
<td>(N_{obs})</td>
<td>780</td>
<td>588</td>
<td>560</td>
<td>592</td>
</tr>
<tr>
<td></td>
<td>(LL)</td>
<td>-1846.17</td>
<td>-1506.76</td>
<td>-1227.25</td>
<td>-1109.84</td>
</tr>
<tr>
<td></td>
<td>(MAPE)</td>
<td>26.49%</td>
<td>27.56%</td>
<td>26.10%</td>
<td>23.63%</td>
</tr>
<tr>
<td></td>
<td>(P_{95})</td>
<td>81.79%</td>
<td>85.71%</td>
<td>80.36%</td>
<td>73.82%</td>
</tr>
<tr>
<td></td>
<td>(P_{99})</td>
<td>88.08%</td>
<td>92.69%</td>
<td>88.04%</td>
<td>80.91%</td>
</tr>
<tr>
<td>Thursday</td>
<td>(N_{obs})</td>
<td>893</td>
<td>536</td>
<td>501</td>
<td>532</td>
</tr>
<tr>
<td></td>
<td>(LL)</td>
<td>-2250.23</td>
<td>-1290.81</td>
<td>-1069.04</td>
<td>-1007.69</td>
</tr>
<tr>
<td></td>
<td>(MAPE)</td>
<td>28.45%</td>
<td>28.62%</td>
<td>24.48%</td>
<td>25.15%</td>
</tr>
<tr>
<td></td>
<td>(P_{95})</td>
<td>84.88%</td>
<td>83.02%</td>
<td>80.04%</td>
<td>77.26%</td>
</tr>
<tr>
<td></td>
<td>(P_{99})</td>
<td>91.15%</td>
<td>88.81%</td>
<td>87.03%</td>
<td>83.27%</td>
</tr>
<tr>
<td>Friday</td>
<td>(N_{obs})</td>
<td>759</td>
<td>420</td>
<td>550</td>
<td>537</td>
</tr>
<tr>
<td></td>
<td>(LL)</td>
<td>-1737.08</td>
<td>-981.07</td>
<td>-1343.89</td>
<td>-988.12</td>
</tr>
<tr>
<td></td>
<td>(MAPE)</td>
<td>25.12%</td>
<td>27.40%</td>
<td>26.61%</td>
<td>24.73%</td>
</tr>
<tr>
<td></td>
<td>(P_{95})</td>
<td>86.96%</td>
<td>82.14%</td>
<td>85.64%</td>
<td>76.54%</td>
</tr>
<tr>
<td></td>
<td>(P_{99})</td>
<td>92.36%</td>
<td>87.38%</td>
<td>90.91%</td>
<td>83.24%</td>
</tr>
<tr>
<td>Saturday</td>
<td>(N_{obs})</td>
<td>308</td>
<td>460</td>
<td>544</td>
<td>498</td>
</tr>
<tr>
<td></td>
<td>(LL)</td>
<td>-381.53</td>
<td>-930.17</td>
<td>-1310.06</td>
<td>-861.93</td>
</tr>
<tr>
<td></td>
<td>(MAPE)</td>
<td>20.78%</td>
<td>27.01%</td>
<td>26.86%</td>
<td>24.22%</td>
</tr>
<tr>
<td></td>
<td>(P_{95})</td>
<td>76.30%</td>
<td>77.17%</td>
<td>76.29%</td>
<td>72.69%</td>
</tr>
<tr>
<td></td>
<td>(P_{99})</td>
<td>80.52%</td>
<td>83.26%</td>
<td>81.43%</td>
<td>76.31%</td>
</tr>
<tr>
<td>Sunday</td>
<td>(N_{obs})</td>
<td>231</td>
<td>318</td>
<td>468</td>
<td>424</td>
</tr>
<tr>
<td></td>
<td>(LL)</td>
<td>-294.75</td>
<td>-657.87</td>
<td>-915.05</td>
<td>-733.15</td>
</tr>
<tr>
<td></td>
<td>(MAPE)</td>
<td>22.98%</td>
<td>26.36%</td>
<td>24.57%</td>
<td>25.61%</td>
</tr>
<tr>
<td></td>
<td>(P_{95})</td>
<td>68.83%</td>
<td>78.93%</td>
<td>75.00%</td>
<td>72.64%</td>
</tr>
<tr>
<td></td>
<td>(P_{99})</td>
<td>75.76%</td>
<td>85.85%</td>
<td>79.49%</td>
<td>78.07%</td>
</tr>
</tbody>
</table>

\(N_{obs}\) is the total usable observations.
Modeling Taxi Demand and Supply in New York City Using Large-Scale Taxi GPS Data

Ci Yang, M.S.
Graduate Research Assistant
Department of Civil and Environmental Engineering
Rutgers, The State University of New Jersey
96 Frelinghuysen Road, Piscataway, NJ 08854
Tel: (631)880-1028
Email: jessie_yang06@yahoo.com

Eric J. Gonzales, Ph.D.
Assistant Professor
Department of Civil and Environmental Engineering
University of Massachusetts Amherst
130 Natural Resources Road, Amherst, MA 01003
Tel: (413)545-0685
Fax: (413)545-9569
Email: gonzales@umass.edu
ABSTRACT

Data from taxicabs equipped with Global Position Systems (GPS) are collected by many transportation agencies, including the Taxi and Limousine Commission in New York City. The raw data sets are too large and complex to analyze directly with many conventional tools, but when the big data are appropriately processed and integrated with Geographic Information Systems (GIS), sophisticated demand models and visualizations of vehicle movements can be developed. These models are useful for providing insights about the nature of travel demand as well as the performance of the street network and the fleet of vehicles that use it. This paper demonstrates how big data collected from GPS in taxicabs can be used to model taxi demand and supply, using 10 months of taxi trip records from New York City. The resulting count models are used to identify locations and times of day when there is a mismatch between the availability of taxicabs and the demand for taxi service in the city. The findings are useful for making decisions about how to regulate and manage the fleet of taxicabs and other transportation systems in New York City.

Keywords: big data, taxi demand modeling, taxi GPS data, transit accessibility, count regression model
INTRODUCTION

Spatially referenced big data provides opportunities to obtain new and useful insights on transportation markets in large urban areas. One such source is the set of trip records that are collected and logged using in-vehicle Global Positioning Systems (GPS) in taxicab fleets. In large cities, tens of thousands of records are collected every day, amounting to data about millions of trips per year. The raw data sets are too large to analyze with conventional tools, and the insights that are gained from looking at descriptive statistics or visualizations of individual vehicle trajectories are limited. A great opportunity exists to improve our understanding of transportation in cities and the specific role of the taxicab market within the transportation system by processing and integrating the data with a Geographic Information System (GIS). Moving beyond simple descriptions and categorizations of the taxi trip data, the development of sophisticated models and visualizations of vehicle movements and demand patterns can provide insights about the nature of urban travel demand, the performance of the street network, and operation of the taxicab fleet that uses it.

Taxicabs are an important mode of public transportation in many urban areas, providing service in the form of a personalized curb-to-curb trip. At times, taxicabs compete with public transit systems include bus, light rail, subway, and commuter trains. At other times, taxis complement transit by carrying passengers from a transit station to their final destination – serving the so-called “last mile.” In the United States, the largest fleet of taxis is operated in New York City (NYC), where yellow medallion taxicabs generated approximately $1.8 billion revenue carrying 240 million passengers in 2005 (Schaller, 2006). All taxicabs in NYC are regulated by the Taxi and Limousine Commission (TLC), which issues medallions and sets the fare structure. As of 2014, there are 13,437 medallions for licensed taxicabs in NYC (Bloomberg and Yassky, 2014), which provide service within the 5 boroughs but focus primarily on serving demand in Manhattan and at the city’s airports, John F. Kennedy International Airport and LaGuardia Airport. Since 2013, a fleet of Street Hail Livery vehicles, known as green cabs, have been issued medallions to serve street hails in northern Manhattan and the outer boroughs, not including the airports.

The TLC requires that all yellow taxicabs are equipped with GPS through the Taxicab Passenger Enhancements Project (TPEP), which records trip data and collects information on fare, payment type, and communicates a trace of the route being traveled to passengers via a
backseat screen. This paper makes use of a detailed set of data that includes records for all 147 million trips served by taxicabs in NYC in the 10-month period from February 1, 2010 to November 28, 2010. Each record includes the date, time, and location of the pick-up and drop-off as well as information about payment, the driver, medallion, and shift. This dataset provides a rich source of information to conduct analysis of the variation of taxi pick-ups and drop-offs across space and time.

In order to effectively plan and manage the fleet of taxicabs, it is necessary to understand what factors drive demand for taxi service, how the use of taxicabs relates to the availability of public transit, and how these patterns vary across different locations in the city and at different times of day. A trip generation model that relates taxi demand to observable characteristics of a neighborhood (e.g., demographics, employment, and transit accessibility) is developed with high temporal and spatial resolution. This paper demonstrates how GPS data from a large set of taxicab data can be used to model demand and supply and how these models can be used to identify locations and times of day when there is a mismatch between the availability of taxicabs and the demand for taxi service. The models are useful for making decisions about how to manage the transportation systems, including the fleet of taxicabs themselves.

Recent work has been done to identify the factors that influence demand for taxicabs within each census tract in NYC based on observable characteristics of each census tract (Yang and Gonzales, 2014). The separate models were developed to estimate the number of taxicab pick-ups and drop-offs within each census tract during each hour of the day, and six important explanatory variables were identified: population, education (percent of population with at least a bachelor’s degree), median age, median income per capita, employment by industry sector, and transit accessibility. Yang and Gonzales (2014) specifically developed a technique to measure and map transit accessibility based on the time that it takes to walk to a transit station and wait to board the next departing vehicle. By modeling taxi demand based on spatially specific information about population characteristics, economic activities as indicated by employment, and the availability of public transit services, the models showed how the influence of various relevant factors changes over different times during the day.

This paper builds on existing working by introducing a novel method to quantify the supply of available taxicabs in a neighborhood based on where passengers are dropped off and the vehicle become available for hire. Although the total supply of taxicabs is itself of interest to
policymakers and regulators, the spatial distribution of this supply has a big effect on where customers are able to hail taxicabs on the street and how long they can expect to wait for an available vehicle. Thus, accounting for the supply of taxis in models of the number of taxicab pick-ups provides additional insights about where taxi demand is being served and where there may be latent or underserved demand for taxicab services. The models that are developed in this paper present additional improvements over previous models by explicitly acknowledging that the number of taxi pick-ups in a census tract is a count process and should be modeled with a count distribution such as a Poisson or negative binomial regression. Both the inclusion of an independent variable for taxicab supply and the use of a count data regression yield detailed models that provide improved insights to the factors that drive taxi demand and affect taxi supply. Furthermore, the visualizations of the modeled results provide greater insights than common techniques that merely plot raw data or show simple aggregations. By developing sophisticated models of supply and demand using the extensive set of data of NYC taxicabs, the underlying patterns in the data reveal how the mode is used and how it may be managed to serve the city better.

LITERATURE REVIEW

There are a number of studies of taxicabs in the literature from the fields of policy and economics. Earlier theoretical models developed for taxicab demand are mainly economic models for the taxicab market (Orr, 1969). Although classic economy theory states that demand and supply will reach equilibrium in a free market, most taxi markets is not actually free, and the roles of regulations that either support or constrain the taxicab industry need to be considered. Furthermore, it has been argued that the price generated by “competitive equilibrium” may be insufficient to cover the social welfare costs (Douglas, 1972) or too high to fully realize certain social benefits (Arnott, 1996). Based on a study of taxicabs in London, England, Beesley (1973) argued that five contributing factors account for the number of taxis per head: 1) independent regulations, 2) the proportion of tourists, 3) income per capita (especially in the center of London), 4) a highly developed radially-oriented railway system, and 5) car ownership. Although these classic papers build a theoretical foundation for modeling taxicab demand, they are based only on aggregated citywide data, such as the medallion price by year (Schreiber,
1975), occurrence of taxicab monopolies by city (Eckert, 1973; Frankena and Pautler, 1984), and the total number of taxicabs by city (Gaunt, 1995).

More recently, attention has been directed toward identifying the factors that influence the generation of taxicab demand. Schaller (1999) developed an empirical time series regression model of NYC to understand the relationship between taxicab revenue per mile and economic activity in the city (measured by employment at eating and drinking places), taxi supply, taxi fare, and bus fare. However, Schaller’s (1999) model is not spatially specific, and is based only on the evolution of citywide totals and averages over time. Other studies compare the supply of taxis in different cities in order to investigate the relationships between taxi demand and factors such as city size, the availability and cost of privately owned cars, the cost of taxi usage, population, and presence of competing modes (Schaller, 2005; Maa, 2005). These studies provide comparisons across different locations, but they do not account for changes with time.

There have been many technology developments that are beneficial to modeling taxicab demand. Examples include in-vehicle Global Positioning Systems (GPS) implanted in modern taxicabs and analytical tools like Geographic Information Systems (GIS), which facilitate analysis of spatially referenced data (Girardin, 2010; Balan 2011; Bai, 2013). As a result, a massive amount of detailed data is recorded automatically for trips served by modern taxicab fleets, such as pick-up locations, drop-off locations, and in some cities a complete track of the route connecting the two (Liang, 2013). These large-scale taxicab data make it possible to build empirical models to understand how taxi trips are generated and distributed across space and time, and how they compete with other transportation modes.

The potential for extracting useful information about taxicab demand and the role of taxis in the broader transportation systems has just begun to be tapped. One recent study considers whether taxicabs operate as a substitute or complement to the public transit system in Boston, Massachusetts (Austin and Zegras, 2012). The study makes use of four days of GPS data from taxicab trips and demographic information about neighborhoods in Boston to develop a Poisson count model for taxicab trip generation.

The model specification is important for properly representing the trip generation process, and since the number of taxicab trips generated per census tract is a count variable, the model should be based on a count distribution. The most common count model is the Poisson model, which has been applied to many fields, such as property and liability insurance (Ismail and
Jemain, 2007), counting organisms in ecology (Hoef and Boveng, 2007), crime incidents (Piza, 2012), and transportation safety (Geedipally, 2007). A critical assumption for the Poisson model is that mean and variance of the response variable are equal. In many cases, the variance of the response variable exceeds the mean, and data with such characteristics are considered over-dispersed. There are several count models that may be suitable for over-dispersed data, such as quasi-Poisson, zero-inflated Poisson, and negative binomial models (Washington et al., 2003).

In this paper, we show how an extensive data set including 10 months of taxicab trip data from NYC can be used to model taxi demand, and the proposed model differs from previous work in four ways: 1) the data set is large, spanning over 2,000 census tracts and including observations from several months; 2) a negative binomial regression is compared to a conventional Poisson model, because a negative binomial distribution is more appropriate for modeling over-dispersed count data; 3) transit accessibility is measured in terms of distance and service headway; and 4) taxicab supply is included as an explanatory variable in order to account for the effect that taxicab availability has on realizing demand.

DATA

The database consists of complete information for all 147 million taxicab trips made in NYC between February 1, 2010 and November 28, 2010. Each record includes spatial and temporal information acquired by GPS (pick-up and drop-off date, time, location), fare (including tolls, tip, total fare paid, and method of payment), and distance travelled. In order to make the large dataset more manageable for regression analysis, the locations of pick-ups and drop-offs are aggregated by NYC census tract, and the times are aggregated by hour of the day.

The response variable is the number of taxicab pick-up counts in each census tract per hour. Six explanatory variables are included, which have been identified as important in a previous study with the same data set (Yang and Gonzales, 2014): population, education (percent of population with at least a bachelor’s degree), median age, median income per capita, employment by industry sector, and transit accessibility. The number of taxicab drop-offs in each census tract during each hour is added as an additional explanatory variable representing the immediately available supply of taxicabs at each location and time. The sources of data for the explanatory factors considered in this study include:

- Drop-off taxi demand per hour aggregated by NYC census tract (DrpOff)
2010 total population that has been aggregated by NYC census tract (Pop)

Median age that has been aggregated by NYC census tract (MedAge)

Percent of education that is higher than bachelors aggregated by NYC census tract (EduBac)

Transit Access Time (TAT), the combined estimated walking time a person must spend to access the nearest station (transit accessibility) and the estimated time that person will wait for transit service (transit level of service);

Total jobs aggregated by NYC census tract (TotJob)

Per capita income aggregated by NYC census tract (CapInc)

Since the model can only be estimated for census tracts with valid data for the response and explanatory variables, the data set is cleaned to eliminate census tracts that do not contain population or employment data. Of the 2,143 census tracts in NYC, 17 census tracts are deleted from the data set.

METHODOLOGY

Linear regression models are inadequate for count data, because the response variable is a count of random events, which cannot be negative. As a result, the models that are developed and compared in this study are count models that are specifically developed to represent count processes. First, the Poisson regression model is introduced, following the reference of Ismail and Jemain (2007). In order to account for the varying effects that each of the explanatory variables have on the response variable at different times of the day, a separate model is estimated for the data in each hour.

Let $Y_i$ be the independent Poisson random variable for the count of taxicab trips in census tract $i = 1 \ldots 2126$. The probability density function of $Y_i$ is defined as:

$$
Pr(Y_i = y_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}
$$

where $y_i$ is the number of observed counts in each census tract $i$, and $\mu_i = E(Y_i) = var(Y_i)$.

The fitted value from a Poisson regression is defined as:
\[ E(Y_i | x_i) = \mu_i = \exp(x_i^T \beta) \]  

(2)

where \( x_i \) is a \( p \times 1 \) vector of the \( p \) explanatory variables considered for each census tract \( i \) and \( \beta \) is a \( p \times 1 \) vector of regression parameters. The Poisson mean on the left side of the definition represents the non-negative expected number of trips generated. Note that for this model form, the logarithm of the observed count varies linearly with the explanatory variables.

The Poisson process is based on the assumption that the mean of the independent random variables is equal to the variance, so Poisson regression is only appropriate if the model count data has equal mean and variance. If the observed variance for the count data exceeds the mean, the data is said to be overdispersed, and an alternative model specification using either quasi-likelihood estimation or negative binomial regression may be used instead. Both methods use a generalized linear model framework. The approach using quasi-likelihood estimation follows Poisson-like assumptions. The mean and variance of the random variable \( Y_i \) are:

\[ E(Y_i) = \mu_i \]  

(3)

\[ Var(Y_i) = \theta \mu_i \]  

(4)

where \( \theta \) is an overdispersion parameter; \( \theta = 1 \) corresponds to the Poisson model. The quasi model formulation leaves the parameters in a natural state and allows standard model diagnostics without losing efficient fitting algorithms.

The negative binomial model is characterized by a quadratic relationship between the variance and mean of the response variable (Hoef, 2007). The density function of negative binomial is defined as in Ismail and Jemain (2007) for the Negative Binomial I model:

\[ Pr(Y_i = y_i) = \frac{\Gamma(y_i + v_i)}{\Gamma(y_i + 1)\Gamma(v_i)} \left( \frac{v_i}{v_i + \mu_i} \right)^{y_i} \left( \frac{\mu_i}{v_i + \mu_i} \right)^{y_i} \]  

(5)

where \( v_i \) is a parameter of the negative binomial distribution that is equivalent to the inverse of the dispersion parameter \( \alpha \). For the negative binomial model, the mean of \( Y_i \) is still \( \mu_i \) as given by (3), but the variance is:

\[ Var(Y_i) = \mu_i + \mu_i^2 v_i^{-1} = \mu_i + \mu_i^2 \alpha \]  

(6)
The mean and the variance will be equal if $\alpha = 0$, so the Poisson distribution is also a special case of the negative binomial distribution. Values of $\alpha > 0$ indicate that the variance exceeds the mean, and the observed distribution is overdispersed.

**Selecting Quasi-Poisson Distribution or Negative Binomial Distribution**

In order to select the most appropriate model specification for the count regression, we must compare the mean and variance of the taxicab pick-up counts, which are the response variable for the proposed model. Table 1 presents a summary by hour of the day of the mean and variance of the total number of taxicab pick-ups per census tract in the 10-month data sample.

<table>
<thead>
<tr>
<th>Hour of day</th>
<th>Mean value of pickup counts</th>
<th>Variance of pickup counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2446</td>
<td>100,804,416</td>
</tr>
<tr>
<td>1</td>
<td>1827</td>
<td>63,070,167</td>
</tr>
<tr>
<td>2</td>
<td>1373</td>
<td>41,751,337</td>
</tr>
<tr>
<td>3</td>
<td>994</td>
<td>24,875,367</td>
</tr>
<tr>
<td>4</td>
<td>711</td>
<td>10,338,265</td>
</tr>
<tr>
<td>5</td>
<td>558</td>
<td>4,547,384</td>
</tr>
<tr>
<td>6</td>
<td>1202</td>
<td>25,984,332</td>
</tr>
<tr>
<td>7</td>
<td>2213</td>
<td>79,611,311</td>
</tr>
<tr>
<td>8</td>
<td>2862</td>
<td>128,381,917</td>
</tr>
<tr>
<td>9</td>
<td>2948</td>
<td>137,735,882</td>
</tr>
<tr>
<td>10</td>
<td>2801</td>
<td>122,508,279</td>
</tr>
<tr>
<td>11</td>
<td>2857</td>
<td>131,214,089</td>
</tr>
<tr>
<td>12</td>
<td>3063</td>
<td>153,234,156</td>
</tr>
<tr>
<td>13</td>
<td>3030</td>
<td>149,228,639</td>
</tr>
<tr>
<td>14</td>
<td>3133</td>
<td>161,516,405</td>
</tr>
<tr>
<td>15</td>
<td>2976</td>
<td>143,190,381</td>
</tr>
<tr>
<td>16</td>
<td>2586</td>
<td>106,107,666</td>
</tr>
<tr>
<td>17</td>
<td>3115</td>
<td>151,429,667</td>
</tr>
<tr>
<td>18</td>
<td>3759</td>
<td>224,925,568</td>
</tr>
<tr>
<td>19</td>
<td>3913</td>
<td>244,880,173</td>
</tr>
<tr>
<td>20</td>
<td>3615</td>
<td>209,472,196</td>
</tr>
<tr>
<td>21</td>
<td>3469</td>
<td>195,600,486</td>
</tr>
<tr>
<td>22</td>
<td>3360</td>
<td>186,065,870</td>
</tr>
<tr>
<td>23</td>
<td>3017</td>
<td>147,131,917</td>
</tr>
<tr>
<td>All_Hours</td>
<td>61,828</td>
<td>57,071,324,807</td>
</tr>
</tbody>
</table>

The variance of taxicab pick-ups per census tract in the 10-month dataset greatly exceeds the mean, as shown in Table 1, which provides an indication that the data is overdispersed. This pattern holds whether all counts from all hours of the day are considered together or the records are broken down by hour of the day. The implication is that the count model for the regression should be appropriate for overdispersed data. To choose between the quasi-Poisson distribution
and the negative binomial distribution, it is necessary to look at how the mean and variance appear to be related. Since a goal of this study is to consider how the effect of explanatory variables changes with the hour of the day, a separate model is estimated for each hour, and the comparison of mean and variance must be considered within each hourly aggregation. Figure 1 presents separate plots comparing count mean and variance for three representative hours: hour 0 is 12:00 A.M. – 1:00 A.M. (midnight); hour 8 is 8:00 A.M. – 9:00 A.M. (morning peak); and hour 17 is 5:00 P.M. – 6:00 P.M. (evening peak).

In order to choose the distribution that most appropriately represents the response variable, the data within each hourly aggregation are divided in 100 subsets using the quantiles of the taxicab pick-up counts. The first category includes taxicab pick-ups for census tracts whose counts fall between the 0 quantile and 0.01 quantile, the second category includes census tract data in the range of the 0.01 quantile and 0.02 quantile, and so on. Within each quantile category, the mean and variance of the included data are calculated, and plotted in Figure 1. A linear function of the form shown in (4) is fitted to estimate $\theta$ and see how well the data matches the assumed relationship for a quasi-Poisson regression model. A quadratic function of the form shown in (6) is fitted to estimate $\alpha$ and see how well the data matches the assumed relationship for a negative binomial regression model. The goodness of fit parameter, $R^2$, is used to identify which specification fits the data better. A value of $R^2$ closer to 1 indicates a better fit. It can be seen in the examples for hours 0, 8, and 17 (Figure 1) that the quadratic function provides a better fit for relating the variance and mean, indicating that the negative binomial distribution is more appropriate for the counts of taxicab pick-ups.

**Selecting Poisson Regression or Negative Binomial Regression**

Although the overdispersed taxicab pick-up data appear to show that a negative binomial regression is a more appropriate model than a Poisson regression, it is also necessary to compare the fit of the models with the explanatory variables that have been identified in the Data section. Several methods can be used to compare the fit of a Poisson regression and a negative binomial regression. Each of the following statistics provides a different type of measure of how well the regression model fits the data set, and these are used to compare the models and select the most appropriate model to relate the explanatory variables to the response variable, the number of taxicab pick-ups per hour.
Figure 1. Plot of the variance vs. mean of the aggregated hourly taxicab pick-up counts for (a) hour 0 (midnight), (b) hour 8 (morning peak), and (c) hour 17 (evening peak). Data are grouped into 100 categories by quantile. The linear equation for the quasi-Poisson model is shown with the blue dotted line. The quadratic equation for the negative binomial model is shown with the red dashed line.
1. Akaike Information Criterion

The Akaike Information Criterion (AIC) is a measure of the relative quality of statistical models by trading off the complexity and goodness of fit. This measure is defined as a function of the Log Likelihood (Fox, 2008; Yan and Su, 2009):

\[ AIC = 2p - 2LL \]  

(8)

where \( p \) is the number of parameters in the model and \( LL \) is the Log Likelihood of the model. A smaller \( AIC \) value represents a better model, and the measure is used to ensure that the model is not overfitted to the data.

2. Goodness-of-Fit Test

The goodness-of-fit test is an analysis of variance (ANOVA) test based on calculating the Pearson’s residuals. The Pearson test statistics is (Cameron and Windmeijer, 1996):

\[ \chi^2 = \sum_{i=1}^{n} e_i^2 \]  

(7)

where the definition of the Pearson residual, \( e_i \), depends on whether the regression is a Poisson model or a negative binomial model:

Poisson Model \[ e_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{\text{var}(Y_i)}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \]  

(8)

Negative Binomial Model \[ e_i = \frac{Y_i - \hat{\mu}_i}{\sqrt{\text{var}(Y_i)}} = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i + \hat{\mu}_i^2a}} \]  

(9)

In both of these cases, \( Y_i \) is the observed count of taxicab pick-ups in census tract \( i \), \( \hat{\mu}_i \) is the modeled count, and there are a total of \( n \) census tracts included in the dataset. The Pearson statistic, \( \chi^2 \), is approximately distributed as chi-square with \( n - p \) degrees of freedom, where \( n \) is the number of observations and \( p \) is the number of predicted parameters (i.e., one parameter per explanatory variable included in the model). If \( \chi^2 > \chi^2_{n-p,0.05} \) or P-value (\( \chi^2 \) test) < 0.05 then the model is statistically different from the observed data at the 95% confidence level. Therefore, we seek a model with P-value (\( \chi^2 \) test) > 0.05.
3. **Sum of Model Deviances**

The sum of squared deviance residuals, $G^2$, is a measure of model fit which is used for Poisson regressions. The sum of model deviances is calculated as (Washington et al., 2003):

$$G^2 = 2 \sum_{i=1}^{n} y_i \ln \frac{y_i}{\hat{\mu}_i}$$  \hspace{1cm} (8)

If the model fit the data perfectly, then $G^2 = 0$, because $\hat{\mu}_i = y_i$ for every census tract $i$. For a count model, such as the Poisson or negative binomial, the observed values are always integers, but the model produces values of $\hat{\mu}_i$ that are continuous, so it is very unlikely to achieve zero sum of squared deviance residuals. Nevertheless, the value of $G^2$ provides a useful measure of the error in the model.

4. **Likelihood Ratio Test**

A second goodness-of-fit test is based on a comparison of the fits of competing models. The likelihood ratio is calculated using the Log Likelihood, $LL$, of each model (Washington et al., 2003).

$$\chi^2_{LL} = -2[LL(\text{model 1}) - LL(\text{model 2})]$$  \hspace{1cm} (7)

where the $\chi^2_{LL}$ statistics is chi-squared distributed with degrees of freedom equal to the difference of the number of parameters estimated in model 1 and model 2. If $\chi^2_{LL}$ is larger than the critical value for the 95% confidence level, then model 1 is said to be statistically different from model 2.

**RESULTS**

Having identified that the negative binomial distribution is more appropriate for the taxicab pick-up data than the quasi-Poisson distribution (as illustrated in Figure 1), it useful to compare the results using a conventional Poisson regression model with the results of negative binomial regression in order to demonstrate the effect of acknowledging the overdispersed response variable. In order to make a comparison between the Poisson and negative binomial regressions, a separate model has been estimated for each hour of the day. The coefficients of the explanatory variables for both models are shown in Table 2.
Yang and Gonzales

Negative Binomial Model Coefficients

Poisson Model Coefficients

Table 2. Parameters of the Poisson and negative binomial regressions for taxicab trip generation
Hour
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Hour
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

DrpOff
0.000048
0.000081
0.000103
0.000194
0.000160
0.000075
0.000031
0.000021
0.000022
0.000025
0.000029
0.000028
0.000028
0.000027
0.000023
0.000023
0.000026
0.000024
0.000023
0.000022
0.000027
0.000030
0.000033
0.000040
DrpOff
0.000311
0.000452
0.000605
0.000861
0.001386
0.001266
0.000177
0.000101
0.000079
0.000070
0.000088
0.000101
0.000106
0.000115
0.000121
0.000138
0.000163
0.000138
0.000110
0.000104
0.000122
0.000138
0.000161
0.000215

Pop
0.000056
0.000026
0.000014
-0.000039
0.000081
0.000162
0.000185
0.000183
0.000175
0.000180
0.000175
0.000146
0.000123
0.000114
0.000102
0.000091
0.000082
0.000073
0.000060
0.000052
0.000039
0.000038
0.000037
0.000041
Pop
0.000088
0.000071
0.000063
0.000042
0.000035*
0.000138
0.000295
0.000310
0.000331
0.000337
0.000309
0.000295
0.000275
0.000258
0.000231
0.000208
0.000188
0.000186
0.000210
0.000215
0.000200
0.000183
0.000152
0.000122

MedAge
0.021619
0.026725
0.021769
0.030535
-0.006056
-0.007020
0.007363
0.013675
0.014264
0.015501
0.015170
0.017751
0.015003
0.021944
0.026618
0.027366
0.027292
0.026853
0.028121
0.029545
0.032460
0.028421
0.027944
0.025840
MedAge
-0.018169
-0.014504*
-0.011550*
-0.011454*
-0.016841
-0.026053
-0.015108*
0.008864*
0.017103*
0.017877*
0.001142*
0.006268*
0.000127*
-0.005119*
-0.007468*
-0.016232*
-0.022523
-0.023651
-0.015490*
-0.010033*
-0.010243*
-0.012317*
-0.003011*
-0.017411*

‘*’ indicates non-significance of the coefficients at 0.05 level

14

EduBac
0.035982
0.031175
0.035571
0.032231
0.039077
0.047770
0.050621
0.054706
0.053651
0.048587
0.045607
0.045852
0.044640
0.046272
0.048343
0.049723
0.050917
0.049711
0.046506
0.046116
0.044167
0.041503
0.038780
0.036323
EduBac
0.037879
0.035677
0.033998
0.031256
0.025647
0.036176
0.058942
0.058380
0.055015
0.058839
0.059663
0.054723
0.051781
0.050155
0.046339
0.042787
0.041864
0.041063
0.044808
0.047140
0.044864
0.044208
0.040828
0.040040

TAT
-0.158105
-0.145743
-0.115985
-0.095762
-0.120689
-0.129105
-0.126587
-0.107053
-0.104637
-0.118816
-0.099687
-0.094150
-0.090213
-0.100170
-0.114502
-0.116026
-0.118006
-0.114158
-0.110181
-0.112516
-0.127821
-0.138917
-0.137978
-0.142699
TAT
-0.071524
-0.076790
-0.084690
-0.082681
-0.076768
-0.081259
-0.080476
-0.082836
-0.085756
-0.088309
-0.075448
-0.080959
-0.076189
-0.075967
-0.075067
-0.072903
-0.069031
-0.071469
-0.073380
-0.076604
-0.074462
-0.070225
-0.071010
-0.069105

TotJob
0.000005
0.000001
-0.000008
-0.000020
-0.000004
0.000005
0.000003
0.000000
-0.000003
-0.000002
0.000004
0.000007
0.000007
0.000006
0.000008
0.000008
0.000008
0.000008
0.000009
0.000009
0.000009
0.000008
0.000006
0.000003
TotJob
0.000045
0.000044
0.000043
0.000039
0.000008*
0.000007*
0.000036
0.000035
0.000035
0.000053
0.000071
0.000082
0.000091
0.000079
0.000076
0.000076
0.000086
0.000090
0.000094
0.000104
0.000099
0.000091
0.000066
0.000055

Proc. NSF Workshop on Big Data and Urban Informatics, Nov. 2015 Page 1163


The six explanatory variables that are significant are the number of drop offs, which represents the available supply of empty taxicabs in a census tract (DrpOff); population (Pop); median age of residents (MedAge); percent of population attaining at least a bachelor’s degree (EduBac); transit access time (TAT); and the total number of jobs located in the census tract (TotJob). Although per capita income had been identified as an important explanatory variable in a previous model based on linear regression (Yang and Gonzales, 2014), the income is highly correlated with the measure of educational attainment (EduBac) in NYC. Therefore, to avoid problems associated with autocorrelation of the explanatory variable, income has been omitted, and the level of education is kept in the models.

The coefficients for all variables in the Poisson regression are statistically significant at level of 0.05 (see top half part of Table 2). While most of the coefficients remain significant when the negative binomial regression is used, median age fails to exhibit significance at the 0.05 level for most hours of the day (see lower part of Table 2). The magnitude of the model parameters is more stable across models for some explanatory variables than others. The results show that regardless of model specification, the taxi supply (DrpOff), education (EduBac), and transit accessibility (TAT) are always significant determinants of taxicab pick-up demand at all times of the day.

In order to determine whether a Poisson regression or a negative binomial regression fits better with the observed hourly taxicab pick-ups, the fours goodness of fit tests introduced previously are used to assess the fit of the two models. These statistics are summarized in Table 3 for each of the 24 Poisson models and 24 negative binomial models (i.e., one for each hour of the day). The results in Table 3 show through many methods of comparison that the negative binomial regression provides a better fit for the data than the Poisson regression. The interpretations of the statistics are as follows:

1) The AIC values for the negative binomial regression models are much lower than for the Poisson regression models.

2) Both models specifications suffer from low p-values for the $\chi^2$ test, so there is substantial variation in the observed data that is not explained by the models. These errors will be reflected in the residuals, so an analysis of the residuals is valuable and necessary.

3) The sums of squared deviances for the negative binomial regression models are much smaller than for the Poisson regression models.
4) The very low p-values of likelihood ratio test statistics suggest that the negative binomial regression model is very different from the Poisson regression model. In light of the numerous differences between the models, the better fit and more appropriate model is the negative binomial regression. That said, the model is not perfect, and although a number of statistically significant explanatory variables and parameters have been identified, these are not sufficient to fully explain the variation in the number of taxicab trips that are generated in census tracts across NYC.

Table 3. Goodness-of-fit statistics for the Poisson (POI) and negative binomial (NB) models

<table>
<thead>
<tr>
<th>Interpretation</th>
<th>AIC</th>
<th>P-value (χ² test)</th>
<th>G²</th>
<th>Likelihood ratio test between POI and NB</th>
</tr>
</thead>
<tbody>
<tr>
<td>hour</td>
<td>POI</td>
<td>NB</td>
<td>POI</td>
<td>NB</td>
</tr>
<tr>
<td>----------------</td>
<td>-----</td>
<td>------------------</td>
<td>----</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>0</td>
<td>4139707</td>
<td>17535</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>3095928</td>
<td>17005</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3196864</td>
<td>16254</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2600073</td>
<td>15684</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2165955</td>
<td>15695</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1398907</td>
<td>15302</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>2866840</td>
<td>16003</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>4510050</td>
<td>16852</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>4949851</td>
<td>16851</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>4703359</td>
<td>16006</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>4188069</td>
<td>15906</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>4357569</td>
<td>15656</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>4609746</td>
<td>15734</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>4620149</td>
<td>15732</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>5142257</td>
<td>16043</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>5072322</td>
<td>16034</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>4449535</td>
<td>15912</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>5147389</td>
<td>16414</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>5958635</td>
<td>16625</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>6493192</td>
<td>16740</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>5795703</td>
<td>16793</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>5569018</td>
<td>17090</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>5591048</td>
<td>17466</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>4965376</td>
<td>17625</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

The interpretation of the model parameters is the same for the negative binomial and Poisson regressions, because both models employ a Log link function. With every unit increase of explanatory variable $x$, the predictor for the response variable increases by a multiplicative factor $\exp(\beta)$. For example the parameter of population in hour 8 of the negative binomial
model is 0.000331, so an increase of population in a census tract by one inhabitant will tend to increase demand by a factor of 1.00033. A positive parameter indicates that the explanatory variable is associated increased numbers of taxicab pick-ups, and a negative parameter indicates an effect of decreased taxicab pick-ups.

In the negative binomial model, the model parameters all have the expected signs. The number of observed taxicab pick-ups increases with taxi supply (DrpOff), population (Pop), education (EduBac), and the total number of jobs within a census tract (TotJob). The effect of transit access time (TAT) is negative, which means that more taxi pick-ups are made in places that have shorter or faster access to subway service. There are a couple possible reasons why a tendency for taxis to be used in the same places that have good transit service should exist. One reason is that people may be likely to take taxis in order to get to or from transit services, so a subway station is a place where a traveler exits the transit system and may look for a taxicab to reach his or her final destination. Another reason is that the types of people and trip purposes that tend to use taxis (e.g., high value of time, unwillingness to search or pay for parking) also tend to occur in parts of the cities that have a lot of transit service. The negative parameter value for TAT is consistent for every hour of the day in the Poisson and negative binomial models, but the precise reason cannot be determined from this regression alone.

One objective of this study is to identify the locations and times of day when there may be a mismatch between taxi demand and supply. One way to investigate this is to look specifically at the Pearson residuals from the models as defined in (9). For a single hour of the day, the residuals for each census tract in the city can be mapped in order to visualize the spatial distribution of the model errors. Maps are presented in Figure 2 for hour 0 (midnight), hour 8 (morning peak), and hour 17 (evening peak), and the color indicates where the model overestimates taxicab pick-ups (i.e., negative residual shown in green) and where the model underestimates taxicab pick-ups (i.e., positive residuals shown in red). The Pearson residual is calculated by dividing the actual residual by the assumed standard deviation, which for the negative binomial model increases as a quadratic function of the mean. This manipulation is used to show the magnitude of error in a standardized manner so that busier census tracts don’t dominate the figure since larger observed and fitted counts will tend to have errors that are larger in magnitude even if those errors are small relative to the expected variance.
Figure 2. Pearson residuals of the negative binomial regression models for (a) Hour 0, (b) Hour 8, and (c) Hour 17.

Taxicab supply is included as an explanatory variable in the model, and the availability of cabs is shown to increase the number of realizing taxicab pick-ups (because the parameter value for DrpOff is positive). A negative residual, which represents an overestimate from the model, provides an indication that there are relatively fewer taxicab pick-ups being demanded relative to the supply of empty cabs available, controlling for the characteristics of the neighborhood. Conversely, a positive residual, which represents an underestimate from the model, provides an indication that there are relatively more taxicab pick-ups being demanded relative to the supply of empty taxicabs. Census tracts that fall into this second condition are of interest, because these
are the locations during each hour of the day, that appear to have insufficient taxicab service relative similar neighborhoods in other parts of the city.

In hour 0 (12:00 A.M. – 1:00 A.M., midnight), the central part of Manhattan and most of the census tracts in the outer boroughs have negative Pearson residuals (colored green in Figure 2a), and the model overestimates the realizing count of taxicab pick-ups. The census tracts with positive Pearson residuals (colored red in Figure 2a), which means that there are more taxicab pick-ups than our model predicts are, are in northern Manhattan, the Lower East Side, western Queens, and the downtown and Williamsburg parts of Brooklyn. These are neighborhoods where there tends to be more night activity than indicated by the explanatory variables and thus more taxi demand. These are the neighborhoods where there is likely to be the largest mismatch between the supply of available taxicabs and the number of people who seek to be picked up by a taxicab.

It is useful to compare the patterns from hour 0 with other hours of the day, because activity patterns in NYC change over the course of the day. In hour 8 (8:00 A.M. – 9:00 A.M., morning peak), the negative Pearson residuals in central Manhattan and much of the outer boroughs reduces in magnitude (yellow or light orange in Figure 2b). This suggests that the binomial regression model provides a better fit during the morning, and taxicab pick-up counts are estimated with less error. One reason for this may be that data associated with the residents of a census tract are most relevant for predicting the number of trips that these residents are likely to make from their homes in the morning. In hour 17 (5:00 P.M. – 6:00 P.M., evening peak), the magnitudes of the Pearson residuals become larger again.

Despite the variations, there are consistent patterns in the maps of the Pearson residuals across all hours of the day. The locations where the model underestimates trips at midnight also tend to have underestimated trips in the morning and evening. Many of these neighborhoods, such as Harlem, the Lower East Side, Astoria, Williamsburg, and Downtown Brooklyn, are dense residential neighborhoods with vibrant local businesses but without the same level of large commercial and tourist activities as are concentrated in much of Manhattan. This may be a reason why these inner neighborhoods are associated with high demand for taxicab pick-ups, but the taxicab fleet has a tendency to focus service in more central parts of the Manhattan. Many of the further outlying neighborhoods in the Bronx, Queens, Brooklyn, and Staten Island tend to have demand overestimated by the model. This is likely because the populations in those areas
either have lower incomes, which make them less likely to choose to pay for taxicab service, or the neighborhood development is at lower densities, which are more conducive to travel by private car than by hailing a taxicab.

The TLC has already changed policies to address some of the mismatch between taxicab supply and demand in NYC. The green Street Hail Livery vehicles (Boro taxi) are allowed to pick passengers only in Manhattan above 96th Street and in the outer boroughs, not including the airports. This coverage area overlaps with many of the underestimated (and potentially underserved) neighborhoods identified in Figure 2. One part of the city that is consistently underestimated in the models but is not within the green cab’s pick-up area is Manhattan’s Lower East Side. One reason for this may be the recent growth that has occurred in the neighborhood, which has increased activity but may not be reflected in the service provided by taxis. Nevertheless, the Lower East Side is an example of a neighborhood area that this modeling approach can identify as being in need of additional taxicab service. These models and figures could be useful tools for transportation planners who want to understand where taxicab service is used, and where more taxicab supply is needed.

CONCLUSION

This study made use of a negative binomial regression model to interpret 10 months of overdispersed taxicab demand data in NYC. Negative binomial regressions have been broadly applied to biology, bio-chemistry, insurance, and finance industries, and this paper shows that the model approach is well suited demand modeling for taxicabs. The raw taxicab dataset includes 147 million records, and in order to make sense of the patterns, the records are aggregated by census tract and hour of the day in order to develop meaningful models of the way that taxicab demand varies across space and time. A number of count models have been considered to model the number of taxicab pick-ups per census tract within an hour of the day, including the Poisson model, quasi-Poisson model, and negative binomial model. By a series of statistical tests, the negative binomial model is shown to be most appropriate for the overdistributed count data. An analysis of the residuals provides useful insights about where taxicab demand appears to be adequately served by the existing supply of taxicabs, and where there is a need for more taxicab service.
The modeling approach started by using important explanatory variables that were identified in a previous modeling effort that used the same dataset (Yang and Gonzales, 2014). An additional explanatory variable was added to represent the taxicab supply, and this is the number of taxicab drop-offs in each census tract during each hour of the day, because each drop-off corresponds to a taxicab becoming available for another customer. The negative binomial regression shows that three explanatory variables are significant during every hour of the day (drop-offs, educational attainment, and transit access time), and two others are significant during most waking hours of the day (population and total number of jobs).

The residual graphs suggest that central Manhattan and most of the outer boroughs have at least enough taxi supply for the demand that is observed, controlling for neighborhood characteristics. The northern part of Manhattan, the Lower East Side, and the western parts of the Queens and Brooklyn all have more observed taxicab pick-ups than the model predicts. The fleet of green Street Hail Livery vehicles serves some of these neighborhoods but not the Lower East Side. The maps of residuals provide some useful insights for the transportation planners to understand when and where we need more taxis.

The taxicab data used to create these models has both spatial and temporal dimensions. The effect of time is accounted for by separating the data by hour of the day, and fitting a negative binomial regression for each hour. The effect of the time of the day that each explanatory variable has on the number of taxicab pick-ups can be observed by comparing the parameter values in Table 2. Additional modeling effort is needed to also account for the spatial correlations in the data set. It is clear from the maps of residuals that adjacent census tracts have correlated performance indicated by the correlated errors. One way to account for these correlations is with a Generalized Linear Mixed Model. Other efforts to improve the model would be to consider additional explanatory variables to account for the activities or popularity of a census tract or to account for the movement of empty taxicabs in search of customers.

Large datasets, such as the records of taxicab trips in NYC, present some challenges, because the raw data is too big to be analyzed directly by conventional methods. By processing the data, and developing models that relate the taxicab data to other sources of information about the characteristics of different parts of the city at different times of day, it is possible to gain useful insights about the role that taxicabs play in the broader transportation system. More
importantly, these insights can be used to plan and improve the transportation system to meet the needs of users.

REFERENCES


Finding Public Transportation Community Structure based on Large-Scale Smart Card Records in Beijing

Song Gao\textsuperscript{a}, Ying Long\textsuperscript{b}

\textsuperscript{a} University of California, Santa Barbara, CA 93106, USA, Email: sgao@geog.ucsb.edu
\textsuperscript{b} Beijing Institute of City Planning, Beijing 100045, China, Email: longying1980@gmail.com

Abstract:

Public transportation in big cities is a crucial part of urban transportation infrastructures. Exploring the spatiotemporal patterns of public trips can help us to understand dynamic transportation patterns and the complex urban systems thus supporting better urban planning and design. The availability of large-scale smart card data (SCD) offers new opportunities to study intra-urban structure and spatial interaction dynamics. In this research, we applied the novel community detection methods from the study of complex networks to examine the dynamic spatial interaction structures of public transportation communities in the Beijing Metropolitan Area. It can help to find the ground-truth community structure of strongly connected traffic analysis zones by public transportation, which may yield insights for urban planners on land use patterns or for transportation engineers on traffic congestion. We also found that the daily community detection results using SCD are different from that using household travel surveys. The SCD results match better with the planned urban area boundary, which means that the actual operation data of public transportation might be a good source to validate the urban planning and development.

Keywords: public transportation, smart card records, spatial interaction, OD flow matrix, community detection, urban big data

1. Introduction

Public transportation in big cities is a crucial part of urban transportation infrastructures. Exploring the spatiotemporal patterns of public trips can help us to understand dynamic transportation patterns and the complex urban systems thus supporting better urban planning and design. The availability of large-scale smart card data (SCD), which is one type of urban Big Data collected from public transportation operations and management institutions, offers new opportunities to study the intra-urban structure and spatial interaction dynamics by zooming...
into individual-based public trips. Previous research has investigated the jobs-housing relationships and commuting patterns using such data and demonstrated comparisons with traditional high-cost travel survey approach (Long et al. 2012, Long & Thill 2013). The study of spatial interactions is one of the traditional researches in Geography and regional science. For regional studies, the functional region is defined by regional geographers based on interactions between its distinctive land-use zones (Johnston et al. 1981). Representation forms of spatial interactions between different zones include human movement, commodity flow, resource allocation, information communication and so on. For the past several decades, studies of spatial interaction processes have mainly been based on the census datasets (Rae 2009; Jang and Yao 2011). Recent fast development in ICT and the availability of big geospatial data (such as mobile phone records, GPS-enabled taxi/cab traces, location-based check-ins) has supported several frontier researches on spatial interactions and networks (Ratti et al. 2010, Gao et al. 2013, Kang et al. 2013, Liu et al. 2014), identifying functional urban regions (Manley 2014), as well as to reveal spatiotemporal intra-urban land use variations from travel patterns (Liu et al. 2012).

In this research, we are interested in extracting origin-destination (OD) flow matrices in the aggregation scale of traffic analysis zones (TAZs) and analyzing the intra-urban spatial interaction patterns revealed by human movements among TAZs using public transportation. Traditional spatial clustering approaches, which group similar spatial objects into classes, are not sufficient to explore the network structure of spatial interactions between different regions. Thus we applied the novel community detection methods from the study of complex networks to examine the dynamic spatial structures of public transportation communities in the Beijing Metropolitan Area ($16,410 \text{ km}^2$). It can help to find the ground-truth community structure of strongly connected TAZs by public transportation, which may yield insights for urban planners on land use patterns or for transportation engineers on traffic congestion.

2. Data

In Beijing, most bus/metro passengers use smart cards when getting on and off buses and metros to pay their fares. Thus, individual OD trips which connect bus stops (or metro stations)
can be extracted directly from the detailed records of SCD. The collected SCD consists of 97.9 million trips from anonymized 10.9 million smart card users during a one-week period from April 5 to April 11, 2010. In order to create the public transportation OD flow matrices in the TAZ level, we first georeferenced all bus stops and metros stations with latitude/longitude coordinates, and then spatially joined them into the total 1911 Beijing TAZ boundaries (see Figure 1). A directed-weighted linkage between two TAZs represents the total number of public trips from the origin-TAZ to the destination-TAZ in a given time interval. Regarding the temporal dynamics, we aggregated the data into different hourly and daily periods to study the spatiotemporal patterns in public transportation, as well as variations between weekdays and weekends.
Figure 1. The study area in administrative districts (different colors) of Beijing and TAZs. Note that the district divisions showed here was the 2010 version without the merges of Xuanwu and Chongwen districts in order to keep consistent with the SCD data collection period.

3. Methodology

In the study of complex networks, a community is defined as a subset (group) of the whole network and the nodes in the same community are densely connected internally and grouped together. The identification of such densely connected nodes in networks is called community detection. Popular community detection methods can be classified into two groups: graph partitioning and hierarchical clustering. Graph partitioning divides a network graph into a set of non-overlapping groups, while hierarchical clustering seeks to build a hierarchy of clusters of nodes, such that for each cluster there are more internal than external connections.

Newman and Girvan (2004) propose a modularity metric to evaluate the quality of a particular division of a network into communities. Modularity compares a proposed division to a null model in which connections between nodes are random. It is defined as the sum of differences between the fraction of edges falling within communities and the expected value of the same quantity under the random null model.

$$ Q = \sum_{k} \sum_{ij \in C} (realflow_{ijk} - estflow_{ijk}) $$

where $k$ is the number of partition communities, $realflow_{ijk}$ gives the actual fraction of interactions between nodes $i$ and $j$ within the same community $C$, and $estflow_{ijk}$ represents the expected values under the random null model or other theoretical models. If the fraction of edges within communities is no better than the null model the modularity $Q=0$, while $Q=1$ indicates the most robust community structure. In practice, modularity values of different real world networks with varying sizes fall into the range 0.3 to 0.7.

We first converted the TAZ-scale OD flow matrices in the consecutive seven days into seven undirected-weighted graphs, where each TAZ can be taken as a node and each OD-flow interaction as a weight edge linking two TAZs. Then, the widely used Newman-modularity-maximization method (Newman 2004) was applied to find the daily public transportation
communities. The modularity measure compares a proposed graph division with a null model in which connections between nodes are random. The modularity was defined as the sum of differences between the fraction of edges falling within communities and the expected value of the same quantity under the null model. In practice, a bottom-up fast greedy algorithm (Clauset et al. 2004, Gao et al. 2013) was adopted for searching an optimized graph partition that maximizes the modularity measure. First, each TAZ started in its own independent cluster of community and the modularity values among all pairs of TAZs for all communities were calculated. Second, a pair of TAZs which has the maximum difference of OD flow compared with the null model should be merged into a community. Third, the modularity of the new graph will be calculated again and then repeating the procedure until the maximum of modularity is found. A larger modularity value indicates a more robust community structure.

4. Results

4.1 Weekdays and Weekends

Table 1 shows the detailed network information of daily community detection results of public transportation OD trips during a week. We find the community consistent pattern in terms of the number of divided groups (6), the average size of each community (202) and the maximum value of modularity (0.457~0.475) in the detection processes. Although the network statistics are similar in the seven days of a week, the spatial distributions of these detected communities lie in slightly different. As shown in Figure 2, the daily community detection results demonstrate that in general geographically cohesive regions that correspond remarkably well with administrative districts in Beijing were identified by weekday public transportation patterns, while some unexpected spatial structures might uncover hidden urban structure that needs further investigation. The suburb public transportation communities usually contain more TAZs than urban central TAZs. There exist strong public transit connections among TAZs which locate along the middle west-to-east corridor including the Chang’ an Avenue in Beijing, where the metro line 1 also runs through the street. Note that the passengers can use smart card when they travelled on metro lines. Surprisingly, most of the southern TAZs in Daxing, Fengtai and Fangshan districts were aggregated into a large transportation community. It indicates there are more frequent intra-public trips within its own community in the southern region than the inter-community trips across other sub-regions of the Metropolitan Area of Beijing. The same giant community pattern lies in the northwest TAZs in Yanqing district and the majority of TAZs in Tongzhou, although there is several connected TAZs from inner districts
to Tongzhou through the Beijing Subway Batong line. Also, it is remarkable to see an enclave in the southern part of Fangshan district has been aggregated into a spatially separated large community north/northeast parts of Beijing (covering a large portion of Chaoyang, Shunyi and Changping districts) in all seven consecutive days, which indicates a strong public transportation connection pattern. The integrated analysis of geographical contexts, land-use types, housing prices, job opportunities, and the prominent points of interest in these regions might offer better explanation about the patterns identified in the community detection results. In addition, a northwest TAZ in Changping district was aggregated into a large number of spatially separated TAZs which belong to inner districts of Beijing (Dongchen, Xicheng, Chongwen and Xuanwu) only in weekends not in weekdays. It reveals a recreation place of interests in the northwest TAZ and attracts a large portion of public travel trips. Potentially, this pattern could help local transportation agency to identify the needs to provide temporal services for increasing public transportation demands in these connected regions.
Figure 2. The spatial distributions of daily community detection results of public transportation OD trips using SCD in a week

Table 1 Daily community detection results of public transportation OD trips in a week

<table>
<thead>
<tr>
<th>Day of Week</th>
<th># of Nodes</th>
<th># of Edges</th>
<th># of Groups</th>
<th>Mean of Community Size</th>
<th>MAX Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>1214</td>
<td>55205</td>
<td>6</td>
<td>202</td>
<td>0.467</td>
</tr>
</tbody>
</table>
We also created an interactive web map for exploring the public transportation community detection results in the geographical context (Figure 3). The online geovisualization of communities for comments and validation using local knowledge can be accessed at http://longy.jimdo.com/data-visualization-1/dv1-10/.

![Image](image.jpg)

**Figure 3.** The interactive web map for exploring the public transportation communities in Beijing

### 4. 2 Comparison with Household Survey Data

Household travel survey is a traditional data-collection approach for acquiring information about residents’ travel behaviors and estimating transportation demands (Beijing Transportation Research Center, 2009). The survey tracks travelers’ socio-economic attributes, as well as trip origin and destination, time and duration, purpose and mode. We applied the same data processing and community detection procedure introduced above to one-day
household survey data. We found that the daily community detection results using the household travel surveys are different from that using SCD.

As shown in Figure 4, there are thirteen communities identified when maximizing the modularity of the TAZ network connected by survey OD trips. It is clear to see that most of the TAZs in suburbs have been grouped into the corresponding outer districts. The community boundaries generally match well with administrative boundaries of Beijing Districts. For those places that didn’t match, especially for the spatial separated communities, it usually indicates some interesting travel patterns, land-use or urban structure, which could be identified through geographical contexts analysis (Gao et al. 2013).

By comparing the community detection results of SCD and household survey data, we also find that the SCD results match better with the Beijing planned urban area boundary (see Figure 5), which means that the actual operation data of publication transportation might be a good source to validate the urban planning and development.
Figure 4. The spatial distribution of community detection results of one-day household survey data.
4.3 Hourly Patterns

Coming back to the SCD, it is interesting to study the spatiotemporal patterns in a micro-time scale. The community detection results of three-hourly aggregated trips, especially the commuting trips at peak hours yield insights on the overall job-related mobility patterns and intra-TAZ spatial interactions using public transportation. Table 2 shows the detailed network information of three-hourly community detection results of SCD during a weekday. We find that the hourly network structures change more (nodes and edges) than that of daily networks. The maximum modularity in hour 18-21 (0.473) and 09-12 (0.470) has the largest values and thus indicates a more robust community structure. For the spatial distribution patterns (see Figure 6), the northern TAZs change more frequently than the southern parts, especially in the Changping District. In addition, similar to the daily patterns, there exist strong public transit connections through the whole day in TAZs that are located along the central west-to-east corridor including the Chang’an Avenue in Beijing, where the Beijing Subway line 1 (west-east) and line 5 (north-south) run through. The interactive web map could also help us to identity underlying patterns by overlaying the detection results on the geographical contexts.
Figure 6. The spatial distributions of three-hourly community detection results of public transportation OD trips using SCD in a weekday

Table 2 Three-hourly community detection results of public transportation OD trips using SCD in a weekday

<table>
<thead>
<tr>
<th>Hours</th>
<th># of Nodes</th>
<th># of Edges</th>
<th># of Groups</th>
<th>Mean of Community Size</th>
<th>MAX Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>06-09</td>
<td>1205</td>
<td>45359</td>
<td>7</td>
<td>172</td>
<td>0.459</td>
</tr>
<tr>
<td>09-12</td>
<td>1207</td>
<td>45607</td>
<td>6</td>
<td>201</td>
<td>0.470</td>
</tr>
<tr>
<td>12-15</td>
<td>1206</td>
<td>55510</td>
<td>6</td>
<td>203</td>
<td>0.462</td>
</tr>
<tr>
<td>15-18</td>
<td>1212</td>
<td>45519</td>
<td>6</td>
<td>202</td>
<td>0.452</td>
</tr>
<tr>
<td>18-21</td>
<td>1208</td>
<td>46378</td>
<td>5</td>
<td>242</td>
<td>0.473</td>
</tr>
<tr>
<td>21-24</td>
<td>1145</td>
<td>31064</td>
<td>7</td>
<td>164</td>
<td>0.451</td>
</tr>
</tbody>
</table>
5 Conclusions and Future Work

In this research, we applied the community detection methods based on the study of complex networks to examine the dynamic spatial interaction structures of public transportation communities in the Beijing Metropolitan Area using SCD. There are several findings based on our experiment results:

First, the community detection results help to identify the functional connected traffic analysis zones by public transportation and most of them are consistent in both weekdays and weekends. Some detected spatially separated TAZs which belong to the same community indicate strong public travel demands in these regions, either for commuting trips on weekdays or for recreational trips on weekends.

Second, the daily community detection results using SCD are different from that using household travel surveys and the SCD community boundary matches better with Beijing urban planned area than the household travel survey.

Third, the hourly network structures change more than that of daily networks; the community detection results also have more variances in spatial distribution.

This research applies a network-analysis approach to investigate the ground-truth community structure of strongly connected TAZs via public transportation, which yields insights urban structure in Beijing from the public transportation functional zone perspective. In further research, we would like to conduct more detailed analysis by integrating land-use data, POI database with human activities from household surveys or social media to give a more holistic view of public transportation using emerging urban big data and computing techniques. In addition, the map matching of these OD trips to actual streets and further analysis could be beneficial for the reliability analysis of street networks and emergency transportation management.
References


Spatially-Explicit Computational Evaluation of Urban Accessibility from a Human’s Point of View

Eran Ben Elia\textsuperscript{1*}, Itzhak Benenson\textsuperscript{1}, Yodan Rofe\textsuperscript{2}, Dmitry Geyzersky\textsuperscript{3}

\textsuperscript{1}Department of Geography and Human Environment, Tel Aviv University, Israel
\textsuperscript{2}J. Blaustein Institutes for Desert Research, Ben Gurion University of the Negev, Israel
\textsuperscript{3}Performit Ltd, Haifa, Israel

eranbe@post.tau.ac.il, bennya@post.tau.ac.il, yrofe@bgu.ac.il, dima@performit.co.il

*corresponding author
Abstract

Accessibility, particularly to public transit is an important consideration in sustainable mobility policies. Various accessibility measures have been suggested in the literature, most at coarse aggregate spatial resolutions of zones or neighborhoods. We suggest accessibility should be measured from the point of view of a human being traversing the transportation network from one building at origin to another at the destination. Spatially-explicit big urban data generated from GIS layers is now widely available. A computational method and application – Access.City - utilizing the power of graph theory and parallel computing are established for calculating access areas based on mode-specific travel times and corresponding paths from origin to destination by car and by public transit, including walking and waiting, at the resolution of individual buildings. The application is tested and applied in a case study involving the evaluation of a new light rail line in the metropolitan area of Tel Aviv and travel to work during the morning peak hour. The results show that unless high spatial resolutions are accounted for in accessibility assessments – biased and even contradictory estimates can arise. Specifically, the contribution of the LRT to accessibility is overrated at low resolutions and for longer journeys. The new method and application can be further employed for developing spatially-explicit equity indices, for investigating the distributional effects of transportation infrastructure investments.

Keywords

Accessibility, equity, high spatial resolution, GIS, graph theory, public transit
Introduction

Accessibility is a concept used in a number of scientific fields such as transport planning, urban planning, economics, and geography. However, accessibility is often misunderstood, poorly defined and poorly measured. In its simplest form accessibility is the ability of people to reach necessary or desired activities using the available transportation modes (Geurs and Ritsema van Eck, 2001; Garb and Levine, 2002; Handy, 1997). Accessibility is regarded as a key criterion to assess the quality of transport policy and land use development (Kenyon et al., 2002; Bristow et al., 2009). Accessibility indicators are used to evaluate the contribution of transportation investments to enable the mobility of people (labor) and goods (products) and hence to an efficient functioning of the economy. Accessibility is also an essential yardstick for evaluating the three common pillars comprising sustainable development: economic development, environmental quality and social equity: (Bruinsma, Nijkamp et al. 1990; Bertolini, 2005; Kwok and Yeh 2004; Feitelson, 2002). The social/environmental justice dimension of sustainable development, in turn, draws the attention towards the distribution of accrued benefits and burdens over different members of society. Accessibility can then be used as an indicator of the extent to which all groups can participate in activities considered ‘normal’ to their society, such as access to employment and essential services (Farrington and Farrington, 2005; Martens, 2012; Lucas, 2012). For all these dimensions, accessibility is thus a key policy indicator, and accessibility measures are a necessary prerequisite for adequate urban planning policy.

While there is no disagreement as to the importance of addressing accessibility in urban planning goals and policy, the main challenge is how to actually measure it. The literature presents a wide range of approaches to measuring accessibility. These vary according to different methods for computing travel times or travel distance, comparison between transportation modes, spatial scale of analysis, network details, software and functions. Due to these differences, accessibility measures have been broadly categorized as either place-based or person-based. Geurs and van Wee (2004) define four types of accessibility measures: (a) Infrastructure-based – service level of transport infrastructure (e.g. “the average travel speed on the road network”); Location-based – accessibility of a location (e.g. “the number of jobs within 30 min. travel from origin locations”); Person-based – individual accessibility (e.g., “the number of activities in which an individual can participate at a given time”; Utility-based – the economic benefits that people derive from access to the destination. A slightly different grouping is provided by Liu and Zhu (2004). Building on their grouping
previous studies can be sorted according to: a) Opportunity-based measures are the number of destinations available within a certain distance/time from origin (e.g. Mavoa et al. 2012; O’Sullivan et al. 2000; Witten et al. 2003; 2011; Ferguson et al. 2013, Martin et al. 2008). b) Gravity-type measures (potential models) refer to the potential of opportunity between two places (e.g. Alam et al. 2010; Minocha et al. 2008; Grengs et al. 2010). c) Utility-based measures – relate accessibility to the notion of consumer surplus and net benefits to different users of the transportation system (e.g. Ben-Akiva and Lerman, 1979). d) Space-time measures – emphasize the range and frequency of the activities in which a person takes part and whether it is possible to sequence them so that all can be undertaken within given space-time constraints (e.g. Neutens et al. 2010; Miller, 1999).

The growing interest in the interdependence between sustainable development and mobility has emphasized the importance of public transit accessibility (Tribby and Zandbergen 2012; Martin et al. 2008; O’Sullivan et al. 2000). Since the disparity of accessibility between cars and public transit provides important information about the degree of car dependence in urban areas, measuring the relative accessibility of transit versus car has been recently analyzed in many urban regions (Ferguson et al. 2013; Grengs et al. 2010; Mao and Nekorchuk, 2013; Mavoa et al. 2012; Blumenberg and Ong, 2001; Hess, 2005; Martin et al. 2008; Kawabata 2009; Salonen and Toivonen 2013). Despite this rather large volume of work developing different kinds of measures, the literature does not provide much guidance on how to choose or apply these measures in policy assessments, in particular, regarding equity.

A key gap exists between how accessibility is primarily addressed in the literature: as a physical-financial construct that can be improved by proper design or costing, and how human beings actually conceive it mentally in their day to day experiences. This point was well summarized by Kwan (1999: 210): “the accessibility experience of individuals in their everyday lives is much more complex than that which can be measured with conventional measures of accessibility”.

A main drawback of almost all previous attempts to measure or model accessibility is fitting the spatial resolution of data availability, processing and subsequent path computations to the scale where real human beings make travel decisions. Thus an adequate view of accessibility demands analysis at a spatial resolution on which a human being normally operates - moving from a point of origin in one building to a destination point in another building, while navigating the transportation network, comprised of different modes, lines and stops. Thus far, this human viewpoint has been difficult to model due to the heavy data
and computation requirements it entails. Consequently, accessibility was mainly evaluated at a coarse and granular scale of municipalities (Ivan et al., 2013); counties, (Karner and Niemeier, 2013); transport analysis zones (Black and Conroy, 1977; Shen, 1998; Bhandari et al., 2009; Ferguson et al., 2013; Foth et al., 2013; Rashidi and Mohammadian, 2011; Haas et al., 2008; Burkey, 2012; Lao and Liu, 2009; Grengs et al., 2010) or neighborhoods (Witten et al. 2011). This kind of analysis commonly results in a discontinuity when evaluating two adjacent zones. Only a few recent studies attempted to model accessibility at parcel level data and even then the results were eventually aggregated for analysis (Mavoa et al. 2012; Tribby and Zandbergen, 2012; Welch 2013; Salonen and Toivonen, 2013). While often sufficient for car-based accessibility, aggregate estimates tend to either over- or under-estimate transit accessibility. Moreover, important components of transit accessibility, such as access and egress walking times, and waiting for transfer times, are not usually included in the calculations of total travel time.

Although spatially-explicit high resolution measurements of accessibility are necessary, they raise severe computational problems. A typical metropolitan area with a population of several millions demands the processing of hundreds of thousands origins and destinations, tens of thousands street segments and thousands of transit lines of different kinds (Benenson et al., 2010; 2011). For example, the metropolitan area of Tel Aviv has a population of 2.5 million and over 300 bus lines. Thus such computations involve the processing of huge volumes of raw data. Until recently, attempting such an endeavor seemed impossible, and there was no option but to aggregate. However, recent developments in graph resolution algorithms, as well as the ability to use parallel computing on the cloud, seem to offer a solution to these problems.

In this paper we make use of a new GIS-based computer application “Access.City” designed, architected and developed by the company Performit®. It manages to make blazing fast accessibility calculations at the resolution of individual buildings using the power of parallel computing and the cloud. Building on the GIS-based character of the application, we developed a technique to make precise, in space and in time, estimates of transit-based accessibility, and, hence, we are able to assess the transportation system at every location and for every hour of the day. Access.City implements the ideas proposed by Benenson et al., (2011) and our recent developments (Benenson et. al., 2014).

The rest of the paper is organized in the following manner. Section 2, presents the operational approach we applied to measure accessibility. Section 3 presents the principles of
Access.City. Section 4 presents the application of Access.City to real data for a case study in the metropolitan area of Tel Aviv (Israel) comparing between a before/after the implementation of a light rail (LRT) line. Section 5, concludes and suggests future research directions.

2. Operational approach

Typically, accessibility is a function of all three components of an urban system: a) land-use distribution of jobs and activities, population densities and socio-economic characteristics of people and their spatiotemporal distribution; b) the transportation system with its road configuration, modes of travel, and the time, cost and impedance of travel from and to any place within the metropolitan area; and c) the demand and benefits that individuals obtain from traveling from their homes to different destinations. Notwithstanding, an urban area is a complex system that changes and adapts, reciprocally. Ideally, if the laws of functioning of each component were known – a co-evolutionary urban system could be reproduced where land use and transport adapt themselves almost seamlessly to individual needs and their spatial distributions. However the rate of change and adaptation is quite different for each component in the urban system. Land-uses and buildings change slowest – over a period of years, even decades, in response to demographic and transport infrastructure changes (for this research we can safely assume they are relatively constant). Conversely, individuals adapt quickly to changes in other subsystems – and to a large degree their behavior is not predictable. We thus measure accessibility as representing people’s potential mobility choices, and not as predicting their actual movements in space-time. The transportation system is more sensitive to policy changes than the land-use system, and therefore we concentrate on evaluating the impacts of changes in this subsystem on people’s accessibility potential to fixed land use destinations. Given this, in order to make sensible sustainable development choices, we have no better existing option but to calculate and compare accessibility in different development scenarios. However, as noted we do this at high spatial resolutions.

Our measure of accessibility is itself already relational. We measure the number of destinations of interest that are available to a person, within a reasonable time frame (between 30 minutes to an hour), using different transportation modes (usually private car and transit, but possibly bicycles and pedestrian movement as well).
The application Access.City implements the measures of accessibility proposed in Benenson et al., (2011). These measures are based on the estimate of the travel time between (O)rigin and (D)estination and are defined for a given transportation (M)ode. For example, between: (B)us and private (C)ar:

- **Bus travel time (BTT):**
  \[
  \text{BTT} = \text{Walk time from origin to a stop of Bus #1} + \text{Waiting time of Bus #1} + \text{Travel time of Bus #1} + [\text{Transfer walk time to Bus #2} + \text{Waiting time of Bus #2} + \text{Travel time of Bus #2}] + [\text{Transfer component related to additional buses}] + \text{Walk time from the final stop to destination} (\text{square brackets denote optional components}).
  \]

- **Car travel time (CTT):**
  \[
  \text{CTT} = \text{Walk time from origin to the parking place} + \text{Car in vehicle time} + \text{Walk time from the final parking place to destination}.
  \]

- **Service area and Access area (MTT = Modal Travel Time):**
  - *Access area:* Given origin O, transportation mode M and travel time \(\tau\), let us define Mode Access Area - MAA\(_O(\tau)\) - as the area containing all destinations D that can be reached from O with M during MTT \(\leq \tau\).
  
  - *Service area:* Given destination D, transportation mode M and travel time \(\tau\), let us define Mode Service Area – MSA\(_D(\tau)\) - as the area containing all origins O from which given destination D can be reached during MTT \(\leq \tau\).

Accessibility is a relational notion. Based on the understanding that the car system already provides adequate levels of accessibility throughout a metropolitan area, we follow Benenson et al.,(2011) in their statement that an advantageous transportation system provides transit accessibility that does not deviate too much from that provided by the car system. This approach comes down to an analysis of gaps in car and transit accessibility as noted earlier. Benenson et al. (2011), proposed a general methodology for comparing transit and car accessibility and we build on this approach.

Access.City includes two main measures of accessibility of a given location, calculated as the ratio of the service or access areas estimated for the two different travel modes.

Given an origin O, the *Bus to Car (B/C) Access Areas ratio* is

\[
\text{AA}_O(\tau) = \frac{\text{BAA}_O(\tau)}{\text{CAA}_O(\tau)} \quad (1)
\]
Given the destination $D$, the *Bus to Car (B/C) Service area ratio* is

$$\text{SA}_{D}(\tau) = \frac{\text{BSA}_{D}(\tau)}{\text{CSA}_{D}(\tau)}$$  \hspace{1cm} (2)

Equations (1) and (2) can be easily specified for any particular type $(k)$ of destinations $D_k$ or origins $O_k$ and, further, towards including destinations' and origins' capacities $D_k,\text{Capacity}$, $O_k,\text{Capacity}$ (say, high-tech enterprises with destination capacity defined as a number of jobs, or low cost dwellings with origin capacity defined as number of dwellings). B/C Service Area ratio to destinations of type $k$ can be defined as the ratio of the sums of capacities of the destinations (e.g., the number of low wage jobs) that can be accessed during time $\tau$ with the Bus and Car:

$$\text{AA}_{O,k}(\tau) = \frac{\sum_{D_k} \{D_k,\text{Capacity} \mid D_k \in \text{BAA}_O(\tau)\}}{\sum_{D_k} \{D_k,\text{Capacity} \mid D_k \in \text{CAA}_O(\tau)\}}$$  \hspace{1cm} (3)

Likewise, B/C Service Area ratio for origins of type $k$ can be defined as the ratio of the sums of capacities of the origins (e.g., number of dwellings in low income neighborhoods) that can be serviced during time $\tau$ with the Bus and Car, respectively:

$$\text{SA}_{D,k}(\tau) = \frac{\sum_{D_k} \{O_k,\text{Capacity} \mid O_k \in \text{BSA}_D(\tau)\}}{\sum_{D_k} \{O_k,\text{Capacity} \mid O_k \in \text{CSA}_D(\tau)\}}$$  \hspace{1cm} (4)

The sum of the nominators of (3) – (4) is the overall capacity of the service/access areas estimated for the bus mode and the sum in the denominator is the overall capacity of the service/access areas estimated for the car mode. Capacity of an origin or destination can obviously be defined in different ways, e.g. capacity of employment can be defined in terms of total number of jobs.

### 3. Access.City

Access.City makes it possible to construct the areas within the urban region that can be reached by car or transit within a given travel time threshold. Based on these areas, it is possible to calculate car-based and transit-based accessibility to different types of land use or set of locations. Provided the required data is available, it is then possible to generate general accessibility indices, using equations (1) and (2), and accessibility indices for origins/destinations of particular types accounting for their capacities, using equations (3) and (4).

While existing tools used in transportation practice, such as TRANSCAD, are able to calculate detailed transit-based accessibility, these tools have not been developed for this
purpose. As a consequence, the calculation is extremely time-consuming, and it is thus prohibitively expensive, to generate accurate estimates of transit accessibility for an entire metropolitan area. For this reason, Access.City operates on a novel graph database engine and employs distributed parallel computing algorithms. Furthermore, Access.City uses high performance No-SQL engine for extremely fast flushing of computation results. The software is being deployed in the cloud as SaaS (Software as a Service). A developed web user interface will allow planners to perform most of GIS planning tasks such as changing transit lines details, adding/removing stops, showing/hiding GIS layers and etc. inside one integrated system without the need for expensive GIS tools. The backend leverages a sophisticated analytical engine using cutting edge graph theory technologies for real-time (or near real-time) processing and distributed parallel computing that exploits GeoTool mapping abilities. The results of planning are shown on interactive maps and easily exported to any GIS tool.

Relative to Benenson et al., (2011), we make progress in three aspects: First, calculating accessibility measures at the resolution of individual buildings is a major step towards understanding individual perception and behavior to changes in the urban and interurban transportation network. Second, incorporating household level socio-economic data into accessibility estimates and, based on this, evaluation of equity implications of transportation plans and projects (work in progress). Third, visualizing accessibility calculations and supplying a user interface for Access.City (still in development and see Figure 1).

The typical inputs for an Access.City database include:

- A layer of roads with the attributes sufficient for constructing a network
- A layer of transit stops and a layer of the transit lines; each line is related to the links and junctions of the road network it passes. Each line is related to its stops.
- Layers of additional transit modes, such as railroads and light rail lines
- A table of transit departure and arrival times
- Layers of urban land uses
- Layers of origin/destinations with capacities given: buildings, commercial facilities, offices and industry, parks and leisure. These layers enable estimating accessibility of specific land-uses and origins/destinations, by types and with respect to their capacities.

Outputs of Access.City include a series of maps, at a chosen resolution, of the accessibility for every spatial unit and corresponding tables and charts.

An example of the parameters applied in the user interface is presented in Figure 1:
Figure 1: Example of Access.City GUI with analysis parameters.

Access.City applies graph theory algorithms (based on Neo4J graph database; http://www.neo4j.org/) to estimate the Access Area for car and public transport (PT). First, every building is interpreted as a B-node. The road network topology is directly transformed into directed graph: junctions into nodes, street section into links (two-way segments are translated into two links) and travel time into the impedance. The transit network depends on time tables and thus the transit graph is constructed for a given time interval. The transit network is translated into a directed graph based on quadruples:

\[
<\text{PTLINEID}, \text{TERMINALDEPARTURETIME}, \text{STOPID}, \text{STOPARRIVALTIME}> 
\]

that are interpreted as a graph PT-nodes. Two PT-nodes

\[ \text{N}_1 = <\text{PTLINEID}_1, \text{TERMINALDEPARTURETIME}_1, \text{STOPID}_1, \text{STOPARRIVALTIME}_1> \]

d and

\[ \text{N}_2 = <\text{PTLINEID}_2, \text{TERMINALDEPARTURETIME}_2, \text{STOPID}_2, \text{STOPARRIVALTIME}_2> \]

in a PT graph are connected by the link \(L_{12}\) in two cases:

1. \(\text{PTLINEID}_1 = \text{PTLINEID}_2, \text{TERMINALDEPARTURETIME}_1 = \text{TERMINALDEPARTURETIME}_2\) and the stop \(\text{STOPID}_2\) is the next to \(\text{STOPID}_1\) on the bus line \(\text{PTLINEID}_1\). The impedance of the link \(L_{12}\) is equal to \(\text{STOPARRIVALTIME}_2 - \text{STOPARRIVALTIME}_1\).
2. The walk time $W_{12}$ between $N_1$ and $N_2$ is less that the maximal possible walk time $\text{WALK}_{\text{max}}$ and STOPARRIVALTIME$_2 - \text{STOPARRIVALTIME}_1 < \text{WALK}_{\text{max}} + \text{WAIT}_{\text{max}}$, where $\text{WAIT}_{\text{max}}$ is a maximal waiting time at a stop.

In road graphs, the B-node is connected to the road node and a road node is connected to a B-node if the walk time between them is less than $W_{\text{max}}$.

In a PT graph a B-node is connected to a PT node $N_1$, given the trip start time $T_{\text{start}}$, the walk time between the B-node and a PT-node is less than $\text{WALK}_{\text{max}}$, and $\text{STOPARRIVALTIME}_1 > T_{\text{start}} + \text{WALK}_{\text{max}}$ and $\text{STOPARRIVALTIME}_1 < T_{\text{start}} + \text{WALK}_{\text{max}} + \text{WAIT}_{\text{max}}$.

The link impedance between the B-node and road node is equal to a walk time between them. For a PT-graph it is a travel time between stops or walk time between building and stop plus waiting time to the arriving bus. Figure 2 presents a description of the process of translation of a typical bus journey from origin to destination into a sequence of connected links.

4. Case Study: The accessibility impact of Tel Aviv’s “Red” Light Rail line

We compare the changes in accessibility due to the introduction of a planned Light Rail line connecting the southern suburbs with the eastern suburbs in an arc that passes through the CBD – called the “Red” line. Figure 3 shows the spatial resolution with the entire metropolitan area on the left panel and a zoom in to the core on the right. This resolution corresponds to squares of $60\text{m}^2$. 

![Figure 2: Translation of the Transit network to a graph](image-url)
Figure 3: Spatial resolution at building level (60 m$^2$): Entire metropolitan area (left) and core (right)

Figure 4, shows accessibility areas by car (15-45 min) and bus (45 min) for the morning peak hour.

Figure 4: Access areas by car and bus - Metropolitan total (left), Metropolitan built (center), Zoom on core (right)

From the maps in Figure 4, it is clear that the car access area is essentially much larger than the bus access area which is limited to the urban core for a 45 minute journey. In 45 minutes
any area in the metropolitan region is accessible by car but only areas in the core are accessible by bus within the same time frame.

Figure 5 shows relative accessibility (bus/car) maps for the core area between 7:00 and 7:30 AM.

![Accessibility Maps](image)

**Figure 5: Access area \( \text{AA}_0(\tau) \) computed for jobs in core area (7:00-7:30 AM).**

From Figure 5 we can see how relative accessibility changes over a 30 minute period. Overall Tel Aviv City and the CBD enjoy the highest transit accessibility. Further away accessibility is much lower.

Figure 6 illustrates the importance of calculating accessibility at high spatial resolutions. On the left panel is the relative accessibility at the resolution of buildings aggregated to transportation analysis zones (TAZ) on the right is the same accessibility index calculated originally at TAZ level. It is easy to verify the contradictory results of the two levels of computation. On the left, high resolution results in seamless changes in accessibility between neighboring zones with higher accessibility in the center compared to lower in the suburbs. Conversely, on the right, low resolution results in the familiar patchwork discontinuity in accessibility levels, with the suburbs having higher accessibility relative to the center.
Figure 6: Access Area $AA_0(\tau)$ – Calculation at high and low spatial resolutions (60 min. starting 7:00AM with 1 transfer).

Figure 7 shows the number of accessible jobs due to the introduction of the Red LRT line. Most of the benefit is around the LRT corridor clearly visible in red-orange colors. This change is most noticeable for short trips. For longer trips the impact of the LRT is limited. That is, buildings close to the LRT line enjoy an improvement in accessibility to all other areas even within short trips. However, this benefit dissipates as journey time and distance from the LRT line increase.

Figure 7: Number of accessible jobs with LRT (30-60 min. trip, starting 7:00AM, 1 transfer)

Figure 8, presents the accessibility levels obtained when the Red LRT line is running in the background.
Once more we can see the gaps created when calculations are conducted at TAZ resolutions. The low spatial resolution inflates the accessibility index, especially for the shorter trips. At high resolutions short trips have low relative accessibility by transit compared to car, except for the LRT corridor. Only with longer trips an increase in transit accessibility is visible. This is partly because car accessibility does not increase much anymore, because it has already reached all the metropolitan area, while transit accessibility continues to increase.

5. Conclusions and further research

In this paper we present a computational method for calculating accessibility measures at high spatial resolution based on “Big” urban data. While, as discussed in the literature, the rationale of these indices is not new, our new method shows the wide gaps in assessments of transit accessibility, when the calculation is done at different resolutions. In particular, the familiar patchwork of accessibility discontinuity when evaluating adjacent zones, is eliminated. The new method utilizes the abilities of graph theory and parallel computing to estimate accessibility similar to a human being moving in urban space from one building as origin to another building as the destination. Our focus was on accessibility to jobs. However, the method can be applied to any trip purpose depending on the availability and quality of the data. The case study we present shows how a new LRT line can be evaluated in terms of transit accessibility improvement. This kind of analysis could be easily incorporated in economic assessments and feasibility studies of planned transportation infrastructure.
While it is expected that transportation investments will result in increasing accessibility and participation rates to vital activities, experience shows that the distribution of these outcomes is not equal throughout the system. Further research which is now in progress utilizes the computed indices to evaluate the equity of the transportation system. Given the finite population of accessibility values (e.g. access area ratio) for each building, we can evaluate how equal is the distribution across these buildings and compare this distribution between several scenarios (e.g. with/without LRT). A very useful equity measure in this respect is the Lorenz curve (Lorenz, 1905) and corresponding Gini index (Gini, 1912). Both measures have been applied extensively in the past and are well known in evaluations of income distribution inequality as well as transportation equity (Fridstrom et al., 2001). Recently researchers have been implementing them in equity of accessibility (Delbose, and Currie, 2011; Welch and Mishra; Kaplan et al., 2014). As these studies implemented TAZ level assessments we plan to apply them to high resolution data as well as develop new equity measures. Moreover, the application used in this paper can be used to compute service area ratios for key institutions such as hospitals or universities, thus evaluating which areas can easily reach them in an adequate time frame. This work, now in progress, will be developed for a future publication.
References


Using connectivity for measuring equity in transit provision Original Research Article *Journal of Transport Geography, Volume 37, May 2014, Pages 82-92*

Sigal Kaplan, Dmitrijs Popoks, Carlo Giacomo Prato, Avishai (Avi) Ceder


Developing a Comprehensive US Transit Accessibility Database

Andrew Owen*  David Levinson†

July 15, 2014

Abstract

This paper discusses the development of a national public transit accessibility evaluation framework, focusing on lessons learned, data source evaluation and selection, calculation methodology, and examples of accessibility evaluation results. In both practice and in research, accessibility evaluation remains experimental and methodologically fragmented. This heightens the “first mover” risk for agencies seeking to implement accessibility-based planning practices, as they must select a method which might produce results that can only be interpreted locally. Development of a common baseline accessibility metric could advance the use of accessibility-based planning. The accessibility evaluation framework described here builds on methods developed in earlier project, extended for use on a national scale and at the Census block level. Application on a national scale involves assembling and processing a comprehensive national database of public transit network topology and travel times. This database incorporates the significant computational advancement of calculating accessibility continuously for every minute within a departure time window of interest. Values for contiguous departure time spans can then be averaged or analyzed for variance over time. This significantly increases computational complexity, but provides a very robust representation of the interaction between transit service frequency and accessibility at multiple departure times.

*University of Minnesota, Department of Civil, Environmental, and Geo-Engineering, aowen@umn.edu
†University of Minnesota, Department of Civil, Environmental, and Geo-Engineering, dlevinson@umn.edu, http://nexus.umn.edu
1 Introduction

The Accessibility Observatory at the University of Minnesota is dedicated to supporting the fields of transportation planning and analysis by developing and applying accessibility-based evaluation metrics in a consistent manner across the United States. A key component of the Observatory’s work has been the development of an integrated software framework for a nationwide evaluation of the accessibility provided by public transit systems. This framework builds on methods developed in partnership with the Minnesota Department of Transportation, extended for use on a national scale at the Census block level. Application on a national scale involves assembling and processing a comprehensive national database of public transit network topology and travel times. This database incorporates the significant computational advancement of calculating accessibility continuously for every minute within a departure time window of interest. Values for contiguous departure time spans can then be averaged or analyzed for variance over time. This significantly increases computational complexity, but provides a very robust representation of the interaction between transit service frequency and accessibility at multiple departure times.

The development of a comprehensive and consistent national public transit accessibility database involved three major components. First, appropriate data sources were identified, collected, and aggregated in a single input geodatabase. Second, a travel time calculation methodology was selected which provides a reasonable and useful representation of expected travel times by public transit. Finally, block-level travel times and the resulting accessibility were calculated in a parallelized, scalable cloud computing environment.

The following sections provide a description of the background and context which informed this project; an overview of the project’s motivation, goals, and implementation; and a discussion of lessons learned and future directions for improving the research and practice of accessibility evaluation.

2 Background

2.1 Accessibility

The concept of accessibility combines the simpler concept of mobility with the understanding that travel is driven by a desire to reach destinations. It is important to distinguish between individual accessibility and locational accessibility: the former seeks to characterize the ease with which travelers might reach destinations, subject to constraints of ability, budget and other barriers; the latter examines accessibility as a spatial phenomenon by considering the costs and benefits of the po-
potential trips offered by transportation systems between origins and destinations of interest. Horner (2004) explored this distinction in the literature and notes that individual accessibility measures are generally poor at “producing ... generalized assessments of intraurban structure,” while accessibility measures are more useful for “understanding relationships between transportation and land use.”

Geurs and Van Wee (2004) provide a taxonomy of accessibility measures and draw a similar distinction between locational and utility-based accessibility measures, and additionally identify infrastructure-based measures which focus chiefly on the conditions of a transportation system and only secondarily (if at all) on the origins and destinations served by it. Metrics which indicate congestion or speeds on highway systems fall into this category.

Locational accessibility can be a particularly useful tool for transportation planners because it provides a way to evaluate the properties of transportation systems at a level that is aggregate enough to avoid the vagaries of individual users’ preferences and constrains, but still detailed enough to provide guidance for planning at the city and regional level. It can be especially useful for multi-modal transportation planning because it is able to provide a level playing field for evaluation modes relative to one another Anderson et al. (2012).

2.2 Accessibility Metrics

Many different implementations of locational accessibility measurement are possible. El-Geneidy and Levinson (2006) provide a practical overview of historical and contemporary approaches. Most contemporary implementations can be traced at least back to Hansen (1959), who proposes a gravity-based weighted accessibility metric to measures the “potential of opportunities for interaction.” Weighted accessibility indices perform well in modeling but raise issues of comparability and consistency: the best-performing functions and parameters are generally estimated independently in each study or study area Ingram (1971). Levine et al. (2012) discuss these challenges in depth during an inter-metropolitan comparison of accessibility; they find it necessary to estimate weighting parameters separately for each metropolitan area and then implement a second model to estimate a single shared parameter from the populations of each. Geurs and Van Wee (2004) also note the increased complexity introduces by the cost weighting parameter.

Perhaps the simplest approach to evaluating locational accessibility is discussed by Ingram (1971) as well as by Morris et al. (1979). Cumulative opportunity measures of accessibility employ a binary weighting function where opportunities are included if they are reachable within a travel cost threshold, and excluded otherwise. Accessibility is calculated for specific travel cost thresholds and the results is a simple count of opportunities that are reachable within each threshold. This
approach involves both advantages and disadvantages. Both calculation and interpretation of the accessibility measure are dramatically simplified, but accessibility must be reported separately for each time threshold of interest, and the metric cannot be finely calibrated to account for varying user preferences, values of time, or other parameters.

2.3 Accessibility of Transit Systems

Lei and Church (2010) provide a review of approaches to evaluating the accessibility provided by transit systems. Developments fall into two categories: changes in the techniques used to calculate travel times by transit, and changes in the ways those travel times are employed to calculate accessibility. The chief technical challenge in evaluations of transit accessibility has been travel times. Prior to the mid-2000s, evaluations of accessibility in transit systems generally operated on simplified representations of transit networks. For example, a bus route might be assigned an average speed, a trip frequency, and hours of service. From these, travel times by transit are estimated rather than measured. Polzin et al. (2002), Beimborn et al. (2003), Wu and Hine (2003), and Shen (2006) follow this general approach. More aggregate evaluations of accessibility, such as those by Kawabata (2003, 2009) and Kawabata and Shen (2007), make use of average travel times reported by transit commuters.

The introduction of the general transit feed specification in 2005 (Google, Inc. 2013) made detailed transit schedules more widely available, while increases in generally-available computing power made their use more feasible. Krizek et al. (2009b), Lei and Church (2010), Benenson et al. (2010), Mavoa et al. (2012), Owen and Levinson (2012), and Dill et al. (2013) demonstrate various calculations of transit accessibility using detailed transit schedules.

Despite their technical differences, these studies of transit accessibility are fairly consistent in the selection and use of travel times to calculate accessibility. In almost every case, the accessibility provided by transit is derived from a single travel time value for each origin/destination pair.

Some work has addressed this limitation. Polzin et al. (2002) proposes a “time-of-day-based” evaluation of transit accessibility, and discusses the fact that transit service levels vary throughout the day. However, the ultimate focus is on variation in demand: after calculating accessibility on a simple hypothetical two-route transit network, the results are scaled based on the distribution of passenger trips throughout the day. Mavoa et al. (2012) address the issue of accessibility variation by reporting a transit frequency measure alongside the accessibility value for each analysis zone. However, the accessibility values themselves are based on travel times calculated at a single departure time. Similarly, Dill et al. (2013) include a single-departure-time accessibility variable when modeling transit ridership in addition to nine other variables describing local service levels.
Lei and Church (2010) propose a method for evaluating transit accessibility that is sensitive to travel time variations throughout the day. This approach calculates accessibility by using detailed schedule information to find the minimum travel time in an arbitrary trip departure window. Owen and Levinson (2012) follow a similar approach, guided by the earlier work of Krizek et al. (2009b). While this makes the selection of a departure time less arbitrary, it still makes the assumption that transit users are willing and able to adjust their departure time, within an arbitrary window, in order to achieve this optimal travel time.

Fan et al. (2013) provide the clearest example of how transit accessibility can be evaluated across multiple discrete departure times. Accessibility values are calculated using travel times based on departures at each hour of the day; these are averaged to produce a single accessibility metric which incorporates travel times at multiple departure times.

Anderson et al. (2013) propose a method for implementing a measurement of transit accessibility that captures the way that accessibility fluctuates continuously over time as trips approach and depart. Owen (2013) implements this approach and demonstrates that continuous accessibility metrics can provide a better description of the variation in transit commute mode share than do metrics evaluated at a single or optimal departure time.

3 Project Overview

3.1 Motivation and Goals

In both practice and in research, accessibility evaluation remains experimental and methodologically fragmented: researchers and planners focusing on different geographical areas often implement different techniques, making it difficult to compare accessibility metrics across different locations. This encourages the development and refinement of improved accessibility evaluation techniques, but heightens the “first mover” risk for agencies seeking to implement accessibility-based planning practices, as they must select a method which might produce results that can only be interpreted locally. Development of a common baseline accessibility metric could advance the use of accessibility-based planning in two ways. First, it can provide a stable target for agencies seeking to implement accessibility-based methods in upcoming planning processes. Second, it can provide researchers a frame of reference against which new developments in accessibility evaluation can be evaluated.

In 2012, the Minnesota Department of Transportation (MnDOT) implemented an “Annual Accessibility Measure for the Twin Cities Metropolitan Area” that provides a methodology for cal-
culating accessibility in the Minneapolis–Saint Paul metropolitan area, which provides evaluation methodology for accessibility to jobs by car and transit (Owen and Levinson, 2012). Development phases of this project relied on proprietary and custom transit schedule data formats because GTFS (described below) had not been adopted by local transit operators (Krizek et al., 2007, 2009a).

Simultaneously, the value of consistent, systematic accessibility evaluations across multiple metropolitan areas was demonstrated by the work of Levine et al. (2012), which collected zone-to-zone travel time information from 38 metropolitan planning organizations to implement a cross-metropolitan evaluation of accessibility by car.

The goal of this project, then, is to combine the lessons learned from these earlier works with recent advancements in transit schedule data format and availability to produce a new, comprehensive dataset of accessibility to jobs by transit.

### 3.2 Data Sources

**Transit Schedules**  
Detailed digital transit schedules in a consistent format are a critical component of this system, and the availability of such data is a relatively recent phenomenon. The General Transit Feed Specification (GTFS) (Google, Inc., 2013) was developed by Google, Inc. and Portland TriMet as a way to provide transit schedules for use in traveler routing and information tools.

Though the initial goal of GTFS was to provide a common format for traveler-focused schedule and routing software, it has also become a key resource for research and analysis of transit systems. Jariyasunant et al. (2011) and Delling et al. (2013) describe recent work in algorithmic approaches to calculating travel times on transit networks. Puchalsky et al. (2012) describe how the stop and schedule data contained in GTFS datasets can strengthen regional planning and forecasting processes. Wong (2013) examines how data currently available in GTFS allows enables network- and agency-level analysis of transit systems, while Catala et al. (2011) identifies ways that the GTFS format could be expanded to support additional uses in transit operations and planning. It would be difficult to overstate the importance of the GTFS data format, and its widespread adoption, in enabling consistent analysis methodology across multiple transit operators.

Despite their importance and digital nature, collection of GTFS datasets can be frustratingly inconsistent and error-prone. While the format of GTFS data itself is standardized, there are no standards for the digital publication of the datasets, and practices vary widely across transit operators. A majority of operators (at least among medium and large metropolitan areas) provide GTFS datasets via a direct web site link. However, even among these, variations in URL naming conventions pose challenges for systematic retrieval. Other operators allow GTFS dataset downloads only after users interactively submit a form or agreement. Still others generate GTFS datasets and
provide them directly to Google, Inc. for use in their popular online routing tool, but release them to the public only in response to direct email requests.

These issues are somewhat mitigated by the web site www.gtfs-data-exchange.com, a crowd-sourced archive of GTFS datasets from around the world. However, the crowd-sourced nature of this resource poses its own challenges. Most importantly, it is very difficult, and in some cases impossible, to validate that a GTFS dataset obtained from www.gtfs-data-exchange.com was originally published by the actual transit operator, or that it has not been modified in some way. For this project, schedules downloaded from this web site are used only when they cannot be obtained directly from a transit operator.

Transit travel time calculations include off-vehicle time costs: waiting at stations as well as time spent accessing an initial station, accessing any required transfers, and accessing the destination after disembarking. This requires a detailed representation of pedestrian facilities in order to calculation walk times between origin and destination census blocks and transit stations. OpenStreetMap (OpenStreetMap Foundation, 2013) provides an open-source dataset with sufficient detail for this purpose. Specifically, the pedestrian network is comprised of OpenStreetMap features with the “footway,” “pedestrian,” and “residential” tags.

**Employment**  Data describing the distribution of labor and employment in the region are drawn from the U.S. Census Bureau’s Longitudinal Employer-Household Dynamics program (LEHD). The workplace area characteristic dataset, which is updated annually, provides Census block-level estimates of employee home and work locations.

In general, LEHD is a useful data source for accessibility evaluation because it is updated yearly and is drawn from actual payroll records collected at the state level — in this case, by the Minnesota Department of Employment and Economic Development. However, it is important to recognize the fact that LEHD data is *synthetic*: while it is based on actual payroll records, the published results are created by an algorithm designed to produce data which are statistically similar to the underlying data, and which converge to the same distribution when aggregated. An analysis by Spear (2011) of LEHD data in transportation analysis found LEHD to be a useful source of both home and work location data, but identified shortcomings related to job locations of federal workers. Tilahun and Levinson (2011) demonstrate the use of LEHD data in contemporary transportation research.

### 3.3 Software

All of the major components of this evaluation system are open source. While this was not a specific goal or requirement, experience from earlier projects suggested some important benefits of
using open source tools. First, open source software is often provides greater flexibility in input and output data formats. This is an important consideration when a project involves multiple stages of data transformation and processing, each performed with a separate tool. Second, open source software can be rapidly customized to fit the project needs. In this project, local customizations to OpenTripPlanner provided more efficient parallelization and allowed for better data interoperability. Finally, open source approaches reduce barriers to replication and validation. Because the output of this project is itself a dataset designed for use in research and practice, it is important that all parts of the methodology — including those implemented using existing software — are thoroughly transparent and understandable.

This project makes use of the following major software packages:

- **OpenTripPlanner** (OTP), an open-source platform for multi-modal journey planning and travel time calculation.

- **PostgreSQL**, an open-source SQL database engine.

- **PostGIS**, a PostgreSQL extension that allows efficient storage and querying of spatial data.

Additionally, numerous smaller scripts and tools were developed specifically for this project.

### 3.4 Data Processing and Organization

**Figure 1** illustrates the basic project architecture and workflow, which is described in the following sections.

**Inputs** The project inputs are stored primarily in a single SQL database. PostgreSQL is used along with the PostGIS extensions; this combination allows spatial and non-spatial data in a single database, automated spatial queries (e.g. to select all origins within a given analysis zone), and spatial indexing methods that accelerate these queries. Specifically, this database contains an extract of all OpenStreetMap pedestrian data for North America; the full block, county, and core-based statistical area (CBSA) datasets from the U.S. Census Bureau; all 2011 resident area characteristics (RAC) and workplace area characteristics (WAC) from the LEHD; and spatial bounds information for all collected GTFS datasets (which are stored separately).

**Calculation** Travel time calculation is an “embarrassingly parallel” problem — a popular term among computer scientists for computation scenarios that can be easily decomposed into many
independent repetitions of the same basic task. Given a suitable data architecture, these tasks can then be performed simultaneously, exponentially increasing the overall calculation speed.

In this case, the calculation of travel times from one origin at one departure time follows exactly the same process as for every other origin and every other departure time. Just under 11.1 million Census blocks (2010) comprise the United States; combined with 1,440 minutes in a day this gives almost 16 billion possible space-time origins. The effective number is less, however, because in blocks with no access to transit service only a single departure time is used — transit travel times vary over the day but walking travel times do not.

The core unit of work — calculating travel times from a single origin at a single departure time — is provided by existing OpenTripPlanner capabilities. The parameters and assumptions involved in these calculations are described in Section 3.5. OTP is natively multithreaded and can efficiently parallelize its work across multiple processors. To achieve efficient parallelization without requiring dedicated supercomputing techniques, the total computation workload is divided into “analysis bundles” which include all information necessary to compute a defined chunk of the final data. Each analysis bundle includes origin locations and IDs; destination locations, IDs, and
opportunity (job) counts; and a unified pedestrian-transit network.

The scope of origins included in each bundle is arbitrary; a useful value of 5,000 origins per bundle was found through trial and error. Figure 3 illustrates the division of a single county into analysis zones, each containing no more than 5,000 census block centroids. Too-small bundles erode overall efficiency by increasing the overhead costs of job tracking and data transfer, while too-big bundles suffer reliability issues: errors do occur, and when they do it is preferable to lose a small amount of completed work rather than a large amount.

Figure 2: Dividing a Census geography (Cook County, IL) into analysis zones containing no more than 5,000 origins

Destinations, on the other hand, are selected geographically. Because travel times are by definition not known until the calculations are complete, it is necessary to include in each bundle all destinations which might be reached from any of the included origins within some maximum time threshold. A buffer of 60 km from the border of the origin zone is used, based on 1 hour of travel at an estimated 60 kph upper limit of the average speed of walk + transit trips. Figure 3 illustrates the spatial selection of destinations for a given set of origins.

OTP’s Analyst module provides a graph builder function that combines pedestrian and transit network data from the input database into a single graph, and locally-developed software merges the graph into an analysis bundle with the appropriate origins and destinations. The bundle is queued
Computations take place on a variable number of cloud computing nodes which are temporarily leased while computations are in progress. (Currently, computing nodes are leased from Amazon Web Services (AWS).) Each node is prepared with OTP Analyst software as well as custom software which retrieves available analysis bundles, initiates accessibility calculations, and stores the results.

**Outputs** The processing of each analysis bundle results in a single data file which records accessibility values for each origin in the bundle. For each origin, this includes an accessibility value for each departure time and for each travel threshold. These values are stored individually and unaggregated to facilitate a wide range of possible analyses. Each result file is tagged with the ID of the analysis zone and range of departure times for which it is valued, and then stored in a compressed format in the cloud storage system.

Because analysis typically takes place at the metropolitan level or smaller, it is rarely necessary to have the entire national result dataset available at once. Instead, custom scripts automate the download of relevant data from the cloud storage system.
### 3.5 Accessibility Calculations

**Transit Travel Time**  
Travel times by transit are calculating using OpenTripPlanner (2013), an open-source software package sponsored by Portland’s TriMet. OpenTripPlanner is a graph-based transit routing system which operates on a unified graph including links representing road, pedestrian, and transit facilities and services.

The time cost of travel by transit is comprised of several components. *Initial access time* refers to the time cost of traveling from the origin to a transit stop or station. *Initial wait time* refers to the time spent after reaching the transit station but before the trip departs. *On-vehicle time* refers to time spent on board a transit vehicle. When transfers are involved, *transfer access time* and *transfer wait time* refer to time spent accessing a secondary transit station and waiting there for the connecting trip. Finally, *destination access time* refers to time spend traveling from the final transit station to the destination. All of these components are included in the calculation of transit travel times.

This analysis makes the assumption that all access portions of the trip — initial, transfer(s), and destination — take place by walking at a speed of 1.38 meters/second along designated pedestrian facilities such as sidewalks, trails, etc. On-vehicle travel time is derived directly from published transit timetables, under an assumption of perfect schedule adherence.

An unlimited number of transfers are allowed. This is somewhat unusual among evaluations of transit accessibility. In many cases travel times are limited to trips involving no more than one or two transfers; this is justified by the observation that in most cities a very large majority (often over 90%) of observed transit trips involve no more than two transfers. However, the shortest-path algorithms typically employed in these evaluations are single-constraint algorithms: they are guaranteed to find the shortest path only when given a single constraint (typically, travel time). When the path search tree is pruned based on an additional constraint such as number of transfers (or, in some cases, transfer wait time), these algorithms provide no insurance against a shorter trip, requiring additional transfers, remaining undiscovered in the pruned space. (Korkmaz and Krunz, 2001; Kuipers et al., 2002; OpenTripPlanner, 2013)

Given the realities of transit networks, it likely that cases where (for example) a three-transfer itinerary provides a faster trip than a two-transfer itinerary are relatively rare. However, given the goal of evaluating the full accessibility provided by a transit system rather than simply the accessibility that is likely to be utilized, this analysis prefers the algorithmically correct approach of using travel time as the single routing constraint and leaving the number of transfers unconstrained.

Just as there is no upper limit on the number of vehicle boardings, there is no lower limit either. Transit and walking are considered to effectively be a single mode. The practical implication of
this is that the shortest path by “transit” is not required to include a transit vehicle. This may seem odd at first, but it allows the most consistent application and interpretation of the travel time calculation methodology. For example, the shortest walking path from an origin to a transit station in some cases passes through potential destinations where job opportunities exist. In other cases, the shortest walking path from an origin to a destination might pass through a transit access point which provides no trips which would reduce the origin–destination travel time. In these situations, enforcing a minimum number of transit boardings would artificially inflate the shortest-path travel times. To avoid this unrealistic requirement, the transit travel times used in this analysis are allowed to include times achieved only by walking.

Continuous Accessibility  Transit accessibility to jobs is evaluated using every minute in the day as a potential departure time. Figure 4 illustrates how accessibility varies minute by minute at a single census block during the 7–9 AM period. Accessibility increases as departure times at nearby stops approach, and then drops after trips depart. Deep troughs in the accessibility profile are associated with times with few or no upcoming trip departures at nearby stops, while sustained periods of high accessibility are associated with periods providing frequent departures. Because of these fluctuations, the average accessibility over the 7–9 AM peak period is significantly lower than the maximum accessibility value over the same period.

3.6 Visualization

This project produces highly detailed accessibility datasets, and some level of aggregation is typically needed to produce easily understandable summary maps. Figure 5 and Figure 6 provide an example of block-level accessibility results mapped at a constant geographic and data scale across four major metropolitan areas: Washington, DC; Atlanta, GA; Seattle, WA; and Minneapolis–Saint Paul, MN. In these maps, accessibility for each Census block has been averaged over the 7–9 AM period. The resulting average accessibility value indicates the number of jobs that a resident of each block could expect to be able to reach given a randomly-selected departure time between 7 and 9 AM.

4 Conclusion

With the framework developed in this project, it is possible to evaluate the accessibility provided by public transit in any area where data is available. Within the United States, the only data limitation
is the availability of transit schedules in GTFS format — all other sources are available with full national coverage. Also significantly, all data is public or available under an open license.

While this project adopted a specific accessibility metric (cumulative opportunities to jobs) and a set of parameters for implementing it, the framework itself provides flexibility. The core OpenTripPlanner software can calculate weighted accessibility; using a different destination type is a trivial modification; various travel time calculation parameters can be easily adjusted. While it is hoped that the accessibility data products described here will be useful for both research and practice, the framework can be used to fit a wide variety of specific accessibility evaluation scenarios. Consistency does not have to mean “one size fits all.”

This project also highlights ways that accessibility evaluation for other transportation modes could be improved. In some ways, public transit networks are the most difficult domain in which to perform this level of evaluation. Accessibility evaluations for car travel, for example, can employ the simplification of using average roads speeds to avoid the need to calculate at multiple departure time; network structure also remains constant over the course of a day. Given appropriate data

Figure 4: Continuous transit accessibility for a single Census block
Figure 5: Transit accessibility results: Washington, DC and Atlanta, GA
Figure 6: Transit accessibility results: Seattle, WA and Minneapolis–Saint Paul, MN
sources, accessibility by car could be calculated for the same block-level resolution at a fraction of the computation costs.

However, this highlights a critical uniqueness of the transit case: travel time data (in the form of schedules) is publicly available. Outside of loop detector-based systems on urban highways (whose data format varies across cities and states), there exists virtually no equivalent for car travel. Comprehensive data sources for road and highways speeds are effectively limited to commercial datasets; efforts to implement a similar evaluation for car accessibility will need to confront this reality.

Acknowledgments

The project described in this article was sponsored by the University of Minnesota’s Center for Transportation Studies. Many of the employed tools and methodological approaches were developed during earlier projects sponsored by the Minnesota Department of Transportation.

References


Impact of Subway Accessibility on Taxi Trip Generation in New York City

Dan Wan, (Corresponding Author)
Ph.D. Student
Department of Civil Engineering, City College of New York
160 Convent Avenue, Marshak Building, Suite MR-910
New York, NY, 10031
Tel: (212) 650-8046
Fax: (212) 650-8374
E-mail: dwan00@citymail.cuny.edu

Camille Kamga, Ph.D.
Assistant Professor
Department of Civil Engineering, City College of New York
160 Convent Avenue, Marshak Building, Suite MR-910
New York, NY, 10031
Tel: (212) 650-8087
Fax: (212) 650-8374
E-mail: ccamga@utrc2.org

M. Anil Yazici, Ph.D.
Research Associate
Region-2 University Transportation Research Center
City College of New York
160 Convent Avenue, Marshak Building, Suite MR-910
New York, NY, 10031
Tel: (212) 650-8071
Fax: (212) 650-8374
E-mail: yazici@utrc2.org

KEYWORDS: Taxi Trip Generation, Subway Accessibility, GPS Data
ABSTRACT

As a super-large city relies on public transport, New York City (NYC) has been working to increase public transport ridership. While both Taxi and Subway play important roles in NYC’s transportation system. For passengers of some taxi trips, even though their taxi travel can be substituted by a subway route serving for the same origin and destination, subway may not be attractive enough for them to change travel mode. Using a large taxi GPS dataset and subway service/geospatial data in NYC, this paper studies the factors related to subway accessibility driving taxi trip generation. Taxi trips, the realized door-to-door travel, are first analyzed descriptively with respect of each trip’s optimal corresponding alternative subway route. The analysis is based on hourly numbers of taxi trips, distribution of origins and destination pairs, time saving and walking time needed if subway was chosen instead of taxi. One main finding is that some percentages of taxi trips do not have an alterative subway route to substitute the taxi travels. This reveals potential underserved subway demands. For those taxi trips that can be substituted by subway routes, models are built for each two hours in weekday and weekend. The results give us insights on the relationship between temporal and spatial variation of taxi demand and subway accessibility-related variables at different time of day and in different boroughs.
INTRODUCTION

The subway system in New York City (NYC), which is one of the most extensive public transportation systems in the world by number of stations, offers rail service 24 hours a day in five boroughs and has an average weekday ridership as 5,156,913 (1). Due to its affordable cost, convenience, and good safety record, subway is an attractive mode in urban areas. Especially in Manhattan, the subway system serves major areas and for many taxi trips one could find an alternative subway ride path. Yet, taxicabs play an important transport role in NYC, carrying about 600,000 passengers by making around 485,000 trips daily, which makes up 11% of travel in the city (2). Such high level of taxi ridership reveals that the subway system may not be attractive enough for some passengers who have specific travel needs in particular conditions. Some researches argued that the transit accessibility, weather, and other socioeconomic factors affect taxi pick-ups and drop-offs (3, 4). In terms of spatial and temporal distribution of taxi trips, transit accessibility has potential to affect traveler choices. The relationship between temporal/spatial variation of taxi demand and subway accessibility may provide important insights on demand for door-to-door travel in particular and mode choice in general. Emerging urban datasets enables researchers to investigate such relationships. The current paper focuses on this relationship by determining each taxi trip’s optimal alternative subway route. We use a large taxi GPS dataset and subway service/geospatial data in NYC.

For most of the taxi trips in NYC, there are corresponding alternative subway route serving for the same origins and destinations. In most instances, those who choose to take a taxi need to walk a certain distance because subway is not a door-to-door service. Assuming taxi passengers are full aware of the availability information of taxi and subway, they may choose taxi for lots of different reasons, e.g. trip purpose, income level and time of day. In this paper we do not have detailed information about passengers’ personal reasons. So we use the characteristics of their generated taxi trips and main accessibility features of the available subway service they refused to choose to gain insights about their service demand and mode choice patterns. In order to analyze the temporal and spatial variation of taxi trip generation and the subway accessibility conditions in which it is more likely to drive taxi trip generation. For each taxi trip, the accessibility features of its corresponding alternative subway routes are:

1) Time of Day;
2) Walk time from origin to subway entrance;
3) In station waiting time;
4) On vehicle travel time;
5) Transfer time (if any) needed by subway route;
6) Walk time from subway exit to destination;
7) Average cost for each passenger.

Regression techniques are widely used to explore the relationships between a response variable and one or more explanatory variables with various models and software. In this study, NYC’s taxi GPS data of one month (totally about 14million) is used to build our response variables for different time periods. April is chosen for its mild weather and national holiday absence. A multiple linear regression model is developed to demonstrate the temporal and spatial variation of taxi trip generation with respect to the subway accessibility. Subway accessibility is represented by the aforementioned features of each taxi trip’s corresponding alternative subway route serving the same origin and destination.

The structure of this study is as following: introduction of this research and a literature review of related studies on taxi trip generation are the first two sections; the next section describes data used in this research and the explanatory variables considered; the section of methodology will present a technique to calculate subway accessibility and perform a multiple linear regression model to identify these factors’ influence; finally, the modeling results and conclusions are presented as the last section.

LITERATURE REVIEW

Both subway and taxi play important roles in urban transportation, however there is not much literature details the impact of subway accessibility on taxi trip generation. Similar to other metropolitan areas with high taxi activity, NYC is one of the most populous urban agglomerations in the world. “Taxis are a vital part of NYC’s transportation network, transporting 25% of all fare-paying bus, subway, taxi and for-hire vehicle passengers traveling within Manhattan” (2). Providing quick, convenient, door-to-door service, taxis function as the equivalent of the suburban family car in NYC (2). However in urban areas, whether taxi service has a competitive or complementary relationship with subway, there is no research result.
As the taxi industry has three constituent elements: passengers, drivers and taxi owners, literature on taxicab demand mainly focuses on model passengers’ demand for taxi (5)(6) before the recent increase in data availability of taxicab trip records. A regression model is developed (5) on the number of taxicabs in the United States and it is found that demand for taxicab licenses in a city was most strongly associated with the number of workers commuting by subway, households without vehicles, and the number of airport taxicab trips. This model predicts the number of taxicab licenses rather than taxi trips demand.

With the advent of GPS technologies in taxicabs, many scholars have used taxi trip GPS records to model traffic flows and other aspects of transit systems, such as the role of taxicab in public transportation, and taxicab’s impact on urban economies, society, and transport (7)(8)(9). Liu et al. (10) analyzed taxi driver behavior by using GPS as daily digital traces of 3000 taxis in a metropolitan area. They study cabdrivers’ behavior features, classify cabdrivers, delineate their operation patterns and get the difference between them. In their paper, they focus on cabdrivers and just demonstrated the potential to utilize pervasive data sets to understand taxi passengers’ choice. Yang and Gonzales (4) build taxi trip generation models for transportation analysis zones at different time of day based on vast information on demographics and socioeconomics. In their paper, they point out that taxi trips are more numerous in places where transit is more accessible because of activity density. Nevertheless there is no detailed determination of this relationship.

Lachapelle and Noland (12) pointed out the possibility that in densely populated Manhattan, commuters access transit by car by riding to train/subway stations, then walk at the destination side of the transit trip, so end up walking more on the destination side of the trip. However, there is no further research on this possibility. And no literature about the acceptable distance that commuters would walk rather than riding car. The accessibility to subway is represented by actual walking distances to transit as a measure of the “proximity” variable by Lewis-Workman and Brod.(13) But they did not consider the service frequency. B. W. Alshalalfah and A. S. Shalaby (14) found that, with dense transit route network, walk access distances would be lower in the downtown area than in other parts of the city of Toronto, Canada. These researches indicate that considering walking distance to transit as part of accessibility is rational. With detailed information about the subway schedule, distribution and taxi trip GPS data, we can get more specific characteristics of each taxi trip and their corresponding alterative subway routes.
This paper builds model to analyze the effects of each subway accessibility feature on taxi trip generation in different time and boroughs.

DATA

In order to analyze temporal and spatial distribution of taxi trips and identify the most influential factors of subway accessibility on taxi trip generation, we use NYC subway schedules in format of General Transit Feed Specification (GTFS), subway system information and taxi GPS data provided by NYC Taxi & Limousine Commission (TLC). GTFS subway schedules are used to determine waiting time for trains, on-train travel time, and transfer time if needed for each alternative subway route. Geocoded locations of NYC subway entrances are used to determine subway proximity (walk access distance on both origin and destination sides). The subway system information includes line (route segment), station name, entrance latitude and longitude, route names, entrance type, entry, exit only, vending, staffing, staff hours, ADA, ADA notes, free crossover, north south street, east west street, and corner. Entrance latitude and longitude are used to calculate walk distances to the closest subway entrance and from the closest subway exit for each taxi trip’s corresponding alternative subway route. When do this calculation, we considered the type of entrance such as entry for full time, entry during daytime or exit only.

The taxi trips data recorded by GPS devices installed in taxicabs are from April 1st, 2010 to April 30th, 2010. The whole dataset includes more than 14 million trips. Each taxicab trip record comprises a number of trip characteristics including a unique key for driver, shift number, trip number, date and time of pick-up and drop-off, locations of pick-up and drop-off, passenger count, payment type, fare amount, tolls amount, tip amount, trip distance, trip time, and others. Among the available fields, trip origin and destination (longitude and latitude), the time of pick-up and drop-off, travel distance, total fare and number of passengers for each taxi trip are used in this paper. Using these variables, origin and destination locations are identified for each trip, and then trips are aggregated based on the time of day.

NYC taxi regulations prohibit taxicabs from pre-arranging service. Meanwhile, For-Hire Vehicles (FHV) are prohibited from picking up street hails. Due to the relatively low proportion of commercial land use in areas outside of the Central Business District (CBD) in NYC, there are more yellow taxicabs in duty at the CBD. NYC as a dense urban setting, especially in the CBD doesn’t have a high private car dependency. So in the CBD, residents and visitors depend on
transit, taxicabs and FHVs more than in the other parts of the city. Taxicab trips included in this paper represent realized demand of street hailing, but not the prearrangement of FHVs.

**METHODOLOGY**

Since Yellow taxis in NYC are administered by the Taxi and Limousine Commission and have some restriction of picking up passengers, 94% of Yellow taxi pick-ups occur either in Manhattan or at one of the airports (2). In this study, taxi GPS data is selected based on the first step calculation for each of the 14 million trips happened in 2010. Regarding the subway line coverage, a sample of subway lines is used for the initial analysis. NYC subway lines #1, and the parts of #2, #3 that overlapping #1’s Manhattan segment are selected as a group based on their North-South alignment in Manhattan and L is selected for its East-West alignment in Manhattan and Brooklyn, so that crosstown travel, travels between two boroughs and travels need transfers could be included in the analysis. Another important reason for selecting these two groups of subway lines is that along their directions there is no subway line is very close to them. This makes them easier to be chosen as the optimal subway route for taxi trips along them.

**Determining Optimal Subway Entrance and Exit to Select Taxi Trips**

For each taxi trip, first we calculate the distances (bird-eye distances) from its origin to each subway entrances and the distance from each subway exit to its destination. The closest subway entrance is decided as the “optimal” entrance if its passenger/s chose subway. The closest subway exit to its destination is the “optimal” exit for its alternative subway route. After each taxi trip’s closest entrance and exit are identified, if both of the destination and the origin are on any one of the four selected lines/segments, this particular taxi trip is selected for further analysis. For those subway routes needs more than one transfer, the “optimal” entrance and exit may not be the real optimal ones. However, because the taxi trips have their “optimal” entrances and exits on our selected lines do not need more than one transfer, their “optimal” in our study is real optimal. We consider taxi trips having their optimal entrances and exits on #1/ #2/ #3 line as T123. Those having their optimal entrances and exits on L line as TL. TT is the name for taxi trips have their optimal entrances on #1/ #2/ #3 line and exits on L line or reverse the order. It means that if passengers of TT use subway to finish the same travel, they need to have one transfer at 14st. We consider T123, TL, and TT separately on purpose of distinguishing characteristics of the spatial distribution.
Then taxi trips are categorized with respect to time of day, day of week and characteristics of the subway alternative (e.g. in-vehicle travel time, waiting time etc.). Figure 1 shows the alignment of selected subway lines and the sketch map of the way to determine an alternative subway route.

![Subway Entrances and Exits of Selected Lines and Sketch Map of DEnt, DExi, the Closest Subway Entrance and the Closest Subway Exit](image)

**Figure 1** Subway Entrances and Exits of Selected Lines and Sketch Map of DEnt, DExi, the Closest Subway Entrance and the Closest Subway Exit

**Walking Time**

This analysis is in view of taxi trips and their corresponding alternative subway routes. For each of T123, TL and TT, total distance to/from taxi trip origin/destination from/to its optimal subway entrance/exit are calculated. By using an average walking speed of 1.4 meters per second (84 meters per minute) (15), we get the walking time ($T_{walk}$ (minute)) needed for each corresponding alternative subway route as:

$$T_{walk} = \frac{D_{Ent} + D_{Exi}}{V_w}$$

Where:

$D_{Ent}$=distance from origin to its optimal available subway entrance (meter), Shown in Figure 1
D_{Exi} = distance from optimal subway exit to destination (meter), Shown in Figure 1

V_w = walking speed (meter per minute)

**Saved Time**

Taxis provide door-to-door service, hence the travel time (T_{taxi}) calculated from the taxi GPS data is assumed to be the door-to-door trip travel time. However, subway travel time depends on access time to the station, the level of service (or the service frequency which determines the waiting time), in-vehicle travel time and transfer time (if any) between different subway lines. The GTFS Schedule Data from the Metropolitan Transportation Authority (MTA) based on service frequency allow us to calculate waiting time (T_{wait}), in-vehicle travel time (T_{vehicle}) and transfer times (T_{transfer}). Accordingly, the difference between the time needed for a taxi trip’s subway route and its own travel time is the saved time (T_{saved}(minute)) by using taxi using taxi:

\[
T_{\text{saved}} = (T_{\text{walk}} + T_{\text{wait}} + T_{\text{vehicle}} + T_{\text{transfer}}) - T_{\text{taxi}}
\]

Where:

T_{\text{walk}} = walking time (T_{\text{walk}}) needed for a taxi trip’s according alternative subway route

T_{\text{wait}} = in stop waiting time for train

T_{\text{vehicle}} = in-vehicle travel time for subway

T_{\text{transfer}} = transfer times (if any)

**Willingness to Pay for Saved Time**

Besides the walking time and saved time, the fare difference between each taxi trip and its subway alternative route is discussed in terms of passengers’ willingness to pay for time savings. We have the number of passengers for each taxi trip and its total fare. So we can use the saved time by taking taxi to get the average rate (R_{\text{min}}(dollar/minute/person)) taxi passengers paid for each saved minute.

\[
R_{\text{min}} = \frac{T_{\text{saved}}}{(F_{\text{taxi}}/N_{\text{taxi}} - F_{\text{sub}})}
\]

Where:
\[ T_{\text{saved}} = \text{time difference between the time used by taxi trip and the time needed for its subway route} \]

\[ F_{\text{taxi}} = \text{total fare for this taxi trip} \]

\[ N_{\text{taxi}} = \text{number of passengers for this taxi trip} \]

\[ F_{\text{sub}} = \text{base subway fare} \]

**Model Building**

Multiple linear regression models are developed for every two hour’s T123, TL and TT in weekdays and weekends respectively. Each model is built using \( T_{\text{walk}}, T_{\text{saved}} \) and \( R_{\text{min}} \) as the explanatory variables and number of T123, TL and TT in every two hours as response variable. The suggested model is aimed to identify each subway accessibility factor’s impacts on taxi trip generation. This information can help policy makers and planners understand the relationship of taxis and subway in large urban areas as two main public transportation modes. The linear regression models are built as follows:

\[
Y = \sum_{i=0}^{n} \beta_i X_i + \varepsilon
\]

Where:

\( Y = \text{taxi trips, response variable} \)

\( X_i = i^{th} \text{ independent variable, } X_0 \text{ is the intercept} \)

\( n = \text{number of independent variables} \)

\( \varepsilon = \text{error, the difference between modeled and observed } Y \)

Overall, the paper presents a case study which involves a unique and application-oriented treatment of a large urban dataset for transportation planning and regulation. Considering the increasing number of taxi GPS datasets around the world, the presented methodology can potentially be applied elsewhere and enable comparative analysis between different demographics and geography.
ANALYSIS AND FINDINGS

Hourly Number of Taxi Trip Generation

Based on the methodology presented in the previous section, the hourly numbers of trips are shown in Figure 2.

![Figure 2 Hourly number of taxi trips generated along subway lines](image)

From Figure 2, the generation of T123 has obvious morning and afternoon peak-hours especially in weekdays, while the other two kinds of trips have peak hours in middle night but no daytime peak-hour. As subway has more frequent service during peak hours to meet higher demand, it is possible that because of more human activity along #1, #2 and #3 during these hours there is more travel demand for all modes including taxi, resulting in the peaks for T123. For nighttime peak hours of TL and TT happened with lower subway service frequencies during nighttime, the reason might be, on one hand, activity features around their origins and destinations are different from T123’s, on the other hand, taxi passengers try to avoid long waiting time for trains and their safety consideration.
The Taxi Trips without Alternative Subway Route

When we look at the specific optimal subway entrance and exit for each T123, and TL, some of them have the same subway station as the optimal entrance and the optimal exit. Two examples of this kind of taxi trip are shown in Figure 3, one is T123 and another one is TL. Due to their travel directions which are perpendicular to the subway lines in some extent, there is no alternative subway route serving for the same origin and destination. It means that passengers of those taxi trips can’t make a subway trip to substitute their taxi trips.

![Figure 3](image)

**Figure 3** Examples of Taxi trips having no Available Subway Route to Substitute Taxi Travelling

The hourly percentages of this kind of taxi trip without alternative subway route are shown in Figure 4. TL has higher percentages from 30% to 50% of trips that can’t be substituted by subway route compared to the percentages of T123 are around 10%. With a lower density of subway lines surrounding L line, there are more taxi trips generated without a choice of subway for riders along it. This phenomenon indicates the possibility that there are some subway demands that go underserved because of inadequate coverage of subway service. For further analysis in this paper, these taxi trips are not included.
Figure 4  Hourly Percentage of Taxi Trip Having the Same Subway Stop as the Closest Entrance and the Closest Exit

Saved Time

Figure 5 shows the average saved time for these three types of taxi trips compared to the time needed for subway trips. From the detailed schedule of subway service, we know that for weekends and nighttime, the service frequencies are less than weekday peak hours. Lower service frequencies indicate longer waiting time in stops for riders. So it is expected that taking taxi in weekends and at nighttime would save riders more time than taking subway.

By comparison between Figure 2 and Figure 5, the generation of TL and TT show similar temporal patterns to that of the hourly average saved times by taking taxis. For TL and TT, the more time riders can save by using taxi, the more taxi trips were generated. In Figure 5 the hourly average saved time of T123 has the same trend as TL and TT which is longer in nighttime and weekends, nevertheless the relationship between it and its hourly numbers of trips shown in Figure 2 is not visually detectable.
As explained in previous section, for each taxi trip’s corresponding alternative subway route, locations of origin and destination determines from origin to optimal subway entrance and from optimal subway exit to destination. Figure 6 shows the average walking times needed.

**Figure 5** Hourly Average Time Saved ($T_{\text{saved}}$) by Taking Taxi Rather Than Subway

**Walking Time**

In weekdays and weekends, the source and sink of taxi trips might change, causing different average needed walking times on origin side and destination side. There are slight difference
between weekdays and weekends and all these three kinds of subway routes have longer walk time needed for drop-off side. Calculation the sum of pick-up side’s average and drop-off side’s average for each kind of trip can give us the average total walk time needed as from 5 minutes to 7 minutes. So there may be a certain time length that people can accept to take subway. When it is longer than that length, choosing a taxi would be an easier choice for them.

Regression Models

Models with and without the intercept are estimated for TL and T123. In all of the models with and without the intercept signs of the coefficients of the explanatory variables are not changed. Therefore, the intercept is removed from Table 1. Except the time periods labeled with “#”, most of the regressions are significant (p-value <0.05). For each regression model, the variables marked with “*” are significant. For those significant regression models without significant variables marked have significant intercepts. Those insignificant ones without significant variables are all excluded in this table.

However for TT, all of the regression models are not significant, so they are not shown in this table. When the variables change one unit, the magnitude and the sign of their coefficients indicate how much is the increase (positive sign) or decrease (negative sign) of the dependent variable for each time period. For instance, from 12am to 2am, the coefficient of $T_{\text{saved}}$ indicates that 1 minute’s increase of time saved by taking subway instead of taxi will result in an increase of 0.918 for the average number of taxi trip.
Table 1: The Coefficient of Models for TL and T123 in Weekdays and Weekends

<table>
<thead>
<tr>
<th>Weekdays</th>
<th>Hour</th>
<th>Model fit statistics</th>
<th>Coefficients of Explanatory Variables</th>
<th>Weekdays</th>
<th>Hour</th>
<th>Model fit statistics</th>
<th>Coefficients of Explanatory Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R² AdjR²</td>
<td>Twalk(min) Tsaved(min) Rmax(d/min)</td>
<td></td>
<td></td>
<td>R² AdjR²</td>
<td>Twalk(min) Tsaved(min) Rmax(d/min)</td>
</tr>
<tr>
<td>TL</td>
<td>12am-2am</td>
<td>0.972 0.957</td>
<td>0.818 0.918* 0.134</td>
<td>12am-2am</td>
<td>0.699 0.548</td>
<td>0.661 4.071* 1.113</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2am-4am</td>
<td>0.941 0.912</td>
<td>-0.312 0.737 0.773</td>
<td>#2am-4am</td>
<td>#4am-6am</td>
<td>0.682 0.733</td>
<td>1.018* 0.526 0.009</td>
</tr>
<tr>
<td></td>
<td>4am-6am</td>
<td>0.705 0.558</td>
<td>0.273 0.548 -0.199</td>
<td>4am-6am</td>
<td>0.688 0.591</td>
<td>0.456* -0.083 1.034</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6am-8am</td>
<td>0.802 0.702</td>
<td>0.332 -1.5 0.531*</td>
<td>#6am-8am</td>
<td>0.68 0.519</td>
<td>0.456* -0.083 1.034</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8am-10am</td>
<td>0.983 0.975</td>
<td>0.27 0.334* 0.649*</td>
<td>#8am-10am</td>
<td>0.68 0.519</td>
<td>0.456* -0.083 1.034</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10am-12pm</td>
<td>0.986 0.978</td>
<td>-0.209 0.442 0.716*</td>
<td>#10am-12pm</td>
<td>0.68 0.519</td>
<td>0.456* -0.083 1.034</td>
<td></td>
</tr>
<tr>
<td>T123</td>
<td>12am-2am</td>
<td>0.690 0.536</td>
<td>-22.241 14.084 6.197</td>
<td>12am-2am</td>
<td>0.733 0.599</td>
<td>-46.527* 41.12 17.405</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2am-4am</td>
<td>0.801 0.702</td>
<td>-9.777* 7.846 13.168</td>
<td>#2am-4am</td>
<td>0.671 0.507</td>
<td>-33.109* 27.48* -23.482</td>
<td></td>
</tr>
<tr>
<td></td>
<td>#4am-6am</td>
<td>0.675 0.513</td>
<td>-68.175* 68.193 4.325</td>
<td>#4am-6am</td>
<td>0.605 0.408</td>
<td>-7.012* 3.712 0.505</td>
<td></td>
</tr>
<tr>
<td></td>
<td>#6am-8am</td>
<td>0.65 0.476</td>
<td>-154.721* 145.908* -7.491</td>
<td>6am-8am</td>
<td>0.811 0.717</td>
<td>-17.025* 15.765 13.863</td>
<td></td>
</tr>
<tr>
<td></td>
<td>#8am-10am</td>
<td>0.725 0.588</td>
<td>-127.584* 117.615* -5.65</td>
<td>#8am-10am</td>
<td>0.811 0.717</td>
<td>-17.025* 15.765 13.863</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10am-12pm</td>
<td>0.808 0.712</td>
<td>-99.997* 85.114* 20.89</td>
<td>10am-12pm</td>
<td>0.87 0.804</td>
<td>-182.775* 188.468* 4.189</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12pm-2pm</td>
<td>0.843 0.764</td>
<td>-110.909* 91.352* 11.828</td>
<td>12pm-2pm</td>
<td>0.693 0.543</td>
<td>-187.061* 177.039* 1.781</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2pm-4pm</td>
<td>0.788 0.682</td>
<td>-182.91* 174.871* 12.13</td>
<td>2pm-4pm</td>
<td>0.67 0.504</td>
<td>-127.401 121.193 1.995</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4pm-6pm</td>
<td>0.781 0.671</td>
<td>-261.697* 255.177* 10.362</td>
<td>4pm-6pm</td>
<td>0.741 0.612</td>
<td>-185.497* 189.223* 8.474</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6pm-8pm</td>
<td>0.806 0.71</td>
<td>-187.204* 164.579* -3.508</td>
<td>6pm-8pm</td>
<td>0.743 0.615</td>
<td>-164.553* 153.31* -16.394</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8pm-10pm</td>
<td>0.604 0.406</td>
<td>-65.586* 56.236* 1.934</td>
<td>#8pm-10pm</td>
<td>0.741 0.615</td>
<td>-56.497* 53.276* 7.172</td>
<td></td>
</tr>
<tr>
<td></td>
<td>#10pm-12am</td>
<td></td>
<td></td>
<td>#10pm-12am</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Effect of walking time

The results suggest that Twalk does not have significant impact on TL in weekdays. For weekends, the coefficients of Twalk have positive sign in early morning which indicates as the distances to subway stations are longer more taxi trips are generated. For T123, all Twalk’s coefficients are negative for both weekdays and weekends. This demonstrates that the further away from subway stations there would be less taxi demands along #1 line and segment of #2 and #3 lines in Manhattan. From the absolute values of Twalk’s coefficients for T123 in different time periods, daytime always has larger absolute values than nighttime. This means that at nighttime shorter walking time needed to take subway is more attractive than in daytime.
The effects of walking time presented here shed light on understanding the spatial patterns of taxi trip generation considering geographical distribution of the subway station. As for taxi passengers traveling along L line crossing East River between Manhattan and Brooklyn, except early weekend morning (4:00am-6:00am), they do not consider walking time needed to take subway as a factor driving them to choose taxi. However taxi passengers of T123 in Manhattan are more inclined to take taxi in places close to subway stations, which means short walking time needed to take subway is not an attractive factor for them to take subway. This might due to some other factor which they attach importance to.

Effect of time saving

$T_{\text{saved}}$ has positive coefficient for TL in a few time periods. For most of the time, taxi passengers of TL do not consider saving time as a factor of choosing taxi. Meanwhile $T_{\text{saved}}$ positively affect the generation of T123 in most time periods. For T123, one more minute saved by taxi trip compared to subway route can increase more taxi trips in daytime than in nighttime. It indicates that taxi riders of T123 in Manhattan are more willing to take taxi to save time in daytime. When taxi can save more time for riders compared to taking subway, it is attracting more riders. This interpretation of $T_{\text{saved}}$’s influence shows that saving time by taxi is one of the main reasons for taxi passengers’ choice in Manhattan along #1, #2, #3 lines but not a significant factor for taxi passengers along L line in most of the time.

Effect of fare for saved time

Most of $R_{\text{min}}$’s significant coefficients have positive signs for TL. Especially during weekday daytime, an increase of the money needed for one minute saved will increase the taxi trip generation along L line. But for T123, none of the coefficients of $R_{\text{min}}$ has a significant impact on taxi trip generation. It might because of their higher income level. They value time saving more than money saving.

In addition, the difference between the features of TL and T123 might be the reason that for TT (taxi trips with corresponsive subway route using L and one of #1, #2, #3 lines involving a transfer) no significant regression model can be built.

Based on the above analysis, we can get some clues with respect to taxi’s role in transportation system. For T123 in both weekdays and weekends, taxi plays as a competitive mode for subway.
Taxi trips in Manhattan are generated surrounding subway stations for the possible reason of high density of human activity. Among the factors analyzed in this paper, their passengers are most sensitive to time saving by taking taxi and do not consider expense of taxi as a factor affecting their choice. For TL (taxi trips along L line), for half of the time, taxi demand does not significantly affected by any variables. So taxi is part of the transportation system with demand independent from subway. This might be a support that taxi runs as a complementary mode for subway. For the other time, expense of taxi affects their passengers’ choice most. The income level of passengers along L line might be a reason.

CONCLUSION AND DISCUSSION

Utilizing a large GPS database of yellow taxi, subway schedule in GTFS and subway system information in NYC, this study develops a novel methodology to understand taxi trip generation patterns considering subway accessibility’s impact. For the first time, we study the accessibility of subway for taxi riders based on their taxi trips’ corresponding alternative subway routes serving the same origins and destinations to have an insight into the reasons for taxi riders’ choice. The accessibility is represented by the factors as time of day, day of week, walk time needed to take subway, time saved by taking taxi, money paid for each minute saved by taxi. We selected three subway lines mainly run in Manhattan (#1, #2, #3) and one between Manhattan and Brooklyn (L) to include typical types of taxi trips as crosstown and across different boroughs. Some descriptive explanations are given for interesting findings. At the end multiple linear regressions are built to identify the important explanatory variables and analyze their impact on taxi trip generation.

The hourly taxi trip generation indicates the taxi demand in Manhattan has peak hours which coincide with the peak hours of subway service. For other borough, higher taxi demand is associated with longer time saved by taking taxi in nighttime compared to subway. Since Manhattan has the highest density of subway service among all boroughs, there is about 10% of T123 cannot be substituted by subway routes, which means that these taxi riders do not have a choice of subway even if they would like to. Meanwhile, with a lower subway density, from 30% to 50% of taxi trips along L line connecting Manhattan and Brooklyn are not provided with an available subway service. Moreover, as Manhattan has a much higher density of subway lines than the other areas in NYC, the difference of the average total walking time needed to take
subway between T123 and TL is less than 2 minutes. So there might be a certain range of the total walking time needed to take subway that people can accept, when it is out of that range people will choose door-to-door transportation modes such as taxi.

Regression models for each two hours in weekdays and weekends include three variables as $T_{walk}(\text{minute})$, $T_{saved}(\text{minute})$, and $R_{min}(\text{dollar/minute})$. The result based on taxi trips’ temporal and spatial distribution gives hints to understand the relationship between taxi and subway as competitive in Manhattan and most of the time complementary in Brooklyn. From the coefficients, total walk times needed to take subway and times saved by taking taxi are the most influential factors for taxi riders of T123. For T123, there are more taxi trips in areas close to subway stations and the more time saved by taking taxi the more taxi trips are generated. The intensity of this relationship changes over times in a day and in different day of a week. For taxi trips along L line, their riders like to save time by taking taxi too. However walking distance is not a significant factor. And for half of the time, taxi demand does not affected by any subway accessibility factors.

This analysis is interesting from a taxi trip generation point of view on a novel perspective considering subway accessibility conditions. This information gotten from this paper can be useful for transit planning and taxi service regulation to help providing more extensive transit coverage for latent riders and better taxi service for places lack of public transit. To understand the accessibility of public transit systematically, we can add bus lines to the research in future.
Reference

1. New York City Taxi and Limousine Commission, 2014 Taxicab Fact Book
7. Drew Austin, P. Christopher Zegras, The taxicab as public transportation in Boston. Massachusetts Institute of Technology. 2011

Exploring the Use of ‘Big Data’ for Analyzing the Dynamic Spatial Patterns of Hotel and Entertainment Development

Li Yin¹, Liang Wu², Sam Cole³, and Laiyun Wu⁴

¹Department of Urban and Regional Planning, SUNY-Buffalo, liyin@buffalo.edu
²College of Information Engineering, China University of Geosciences, wuliang133@189.cn
³Urban and Regional Planning, SUNY-Buffalo, samcole@buffalo.edu
⁴Department of Urban and Regional Planning, SUNY-Buffalo, laiyunwu@buffalo.edu

Keyword: Traditional data, ‘big data’, hotel development
Introduction

Tourism cluster consists of linked industry components such as accommodation and retail, which provide a variety of products and services to visitors. The relationship between tourism product similarity and spatial proximity has not been adequately studied empirically (Weidenfeld et al., 2010). The rapid expanding of digital data from social media and other internet sources such as TripAdvisor and Google Map has become a source of ‘big data’ on tourism industry (O’Leary, 2013). Combing social media and commercial data sets with traditional sources improve opportunities for constructing and testing more complex industry models and theories. We are exploring the possibilities of use of these data to analyze co-location synergies.

This study focuses on three coastal tourist cities as our study area including Savannah, GA, Charleston, NC, and New Orleans, LA. Building on previous tourism development modeling and spatial studies (Cole, 2009; and Yin, 2007), this study uses GIS to visualize spatial clustering patterns of accommodation for analyzing the patterns of tourism business agglomeration. We use data mining method to help explore the use of large amounts of big data for hotel and entertainment enterprises from different sources. The technologies involve data discovery, data extraction, and database management. The findings of this study can help to study the dynamics of spatial patterns of hotel and entertainment development. This paper also provides a protocol to combine traditional data sources with ‘big data’ such as social media and other internet data for the study of tourism development.

Method

Traditional Data Collection
Hotel data were collected from ReferenceUSA for a period of 15 years (1997 to 2011), in tabular format. These tables were geocoded in ArcGIS 10.1. Most of the hotels are automatically matched while the unmatched ones were pinned out on the map manually.
Updating Data: Spatial Visualization & Big Data

Data processing and comparison with maps provided by Google and other websites suggested that there are problems with precision and completeness of the hotel data collected from ReferenceUSA such as duplicate and missing enterprises between years, use of postal office boxes instead of physical addresses, etc. Considerable effort is needed to salvage sufficient reliable information before conducting analysis.

Alternative data on hotel and entertainment were collected from Google, Tripadvisor, and hotel or entertainment enterprise’s websites. Both Google and Tripadvisor have maps and other information about hotels and individual entertainment enterprise such as address, rating, price, nearby amenity, and reviews from people who used their services. Individual business’s website also provides information on history of the business, such as when it was established and whether its name was changed.

We followed the data mining process as shown in Figure 1. Upon completion of extracting URLs of all hotels, specific information for each hotel was extract based on the URL. For each hotel, various types of information were extracted, including hotel name, rating, price, address, phone number, reviews, etc. When extracting information, web source of the hotel’s website was analyzed to get the specific data through specific expressions. We extracted comments and hotel review between 2003 and 2014 in the extraction process. After that, the review data was categorized by year and the number of total review and the annual review for each hotel were calculated.
Visualizing Clusters

Kernel density tool in ArcMap 10.1 was used to map out the point density of hotels in gradual color. Kernel density maps are smoothly curved surfaces fitted over each point. The surface value reaches the highest at the point’s location and decreases with the increment of distance from the point, reaching zero at the Search radius distance from the point.

Findings

Updated Hotel Data

Figure 2 shows hotels from the traditional database combined with big data. Black dots are hotels from the original ReferenceUSA database. Red dots are updated hotels using both spatial visualization and big data. In all three cities, there are substantial number of hotels missing from the original database, especially in recent years.
Kernel Density of Hotels

Using both the original and updated hotel database, the density classes of hotels were categorized into four levels. Level 1 stands for a relative low density and level 4 means a relative high density. In Savannah, level 1, 2, 3, 4 mean there are 0-8, 8-11, 11-24, 24 and above hotels per square kilometer respectively. In Savannah, cluster from the updated database along Savannah River were growing year by year, starting from 1997. There were also a cluster growing and heading south in southern downtown area, but later declining, and then growing again since 2006. Another new cluster was detected along highway in the western part. This cluster doesn’t become larger until 2006. The hotel clusters from the original database in Savannah was developed along Savannah River first, and then was attracted by the highways, which located in the west Savannah downtown.
Similar to Savannah, the hotel densities of Charleston were classified into four classes. The four level of density are 0-15, 15-30, 30-50, 50 and above hotels per square kilometer respectively. In Charleston, two old clusters in the north and center were growing and finally merged into one big cluster. There is a new cluster emerged on the southern waterfront since 2004. Generally, hotels in Charleston were growing at a fast speed from 2000 to 2010. The changing pattern of hotel clusters in Charleston from the original database is different from updated ones. There is no new hotel clusters were detected in the original dataset.

In New Orleans, The four level of density are 0-27, 27-49, 49-83, 83 and above hotels per square kilometer respectively. There are three clusters in New Orleans, one in north, located in the French quarter district; one in the center, located around the Canal Street and one in south, just above the highway US-90. The cluster located in French quarter stopped growing since 2002, while the cluster around Canal Street continued growing throughout this 13-year period. However, a new cluster in the southeastern area didn’t appear until 2002. By using kernel density, changing patterns of these clusters can be visualized and compared between the original and updated database. In New Orleans, the hotel patterns are basically the same as updated ones. It’s probably because of the size of sample in New Orleans are significantly larger than the other two samples. The patterns were not influenced too much even if there are missing hotels or mismatched hotels in the original data.
Conclusion

This study explores the possibilities of combining big data and traditional data sources to improve opportunities for analyzing co-location synergies in the hotel industry. The findings suggested that ‘big data’ serves as a useful data source to supplement traditional data to prepare better empirical data for analysis.
Reference:


“The magic’s in the recipe”
Urban Diversity and Popular Amenities*

Daniel Arribas-Bel$^a$ and Jessie Bakens$^{b,c}$

$^a$ School of Geography, Earth and Environmental Sciences, University of Birmingham, United Kingdom
$^b$ Department of Spatial Economics, VU University, Amsterdam, The Netherlands
$^c$ Tinbergen Institute, Amsterdam, The Netherlands

July 15, 2014

**IMPORTANT**: preliminary version; please, do not distribute without permission of the authors.

Abstract

This paper uses a novel source of (big) data to analyze the main factors behind the popularity of urban amenities in The Netherlands. In particular, we collect data from the location-based service Foursquare and employ it to obtain a rich catalogue of restaurant locations, as well as a database of other urban amenities. This, combined with traditional sources of socio-economic data, allows us to estimate regressions at the area and venue levels, uncovering the main determinants of the popularity of specific restaurants as well as of entire areas or neighborhoods of a city. In doing so, we contribute to the existing literature along three main dimensions: we provide insight and new knowledge about urban systems, in particular about the under-studied aspect of urban amenities and the role diversity plays as a determining factor; we demonstrate the use of a novel source of data available to urban researchers as a byproduct to improve the understanding of phenomena of interest not only to researchers but to practitioners such as urban planners and business owners; and we quantify, document and characterize some of the biases inherent to these new sources of data in the context of urban applications. Results point to a robust and consistent association between restaurant diversity and popularity, both at the area, the aggregated restaurant, and individual restaurant level.

**Keywords**: urban amenities, diversity, Foursquare

**JEL-classification**: O18, R00

---

*The authors acknowledge financial support from the NORFACE research program on Migration in Europe – Social, Economic, Cultural and Policy Dynamics. Thanks to participants of the VU-Eureka seminar, TI workshop, ADETRE, and Gabriel Ahlfeldt for very useful discussions of and comments on previous versions of this paper. Any errors remain responsibility of the authors.
1 Introduction

The seminal paper of Glaeser et al. (2001) on the consumer city has spurred the conviction that cities are (also) locations of consumption: as income grows and services gain relevance in the economy, the consumption dimension of a city takes priority over its production one. Glaeser et al. (2001) put forward that, besides good public services and speed of movement, the availability of aesthetically beautiful buildings, the physical setting, and the presence of a rich variety of services and consumer goods are the main determinants of an attractive consumer city.

Although the dual role of cities is widely recognized by the literature, very little research has been devoted to analyze and identify the mechanisms that lead to attractive consumer cities. Most of the ingredients for this mechanism are well understood. The sheer quantity of restaurants, bars and other cultural amenities such as museums or theaters positively affects the attractiveness of a city (Glaeser et al., 2001; Dalmazzo and de Blasio, 2011); the presence of open space in the form of parks or urban forests is also positively valued (see for example Carlino and Saiz, 2008). Even the population composition of a city has been treated as a consumer amenity in itself (Ottaviano and Peri, 2006, 2005): either via a direct channel or through its effect on the supply of ethnic goods, immigrants are considered to bring a different taste to the city from which consumers derive utility.

If the ingredients of an attractive consumer city are clear, little is known about the recipe. How these elements combine into positively valued consumption locations is a far less explored topic. Not all areas (or sometimes only a few areas) of attractive cities are considered as such. There can be many economic mechanisms that influence the spatial arrangement of amenities within a city. Preference for diversity can lead to areas in which very different amenities coexist and exert a positive externality among themselves, as in the Jacobian (for Jane Jacobs) externalities. Equally, agglomeration externalities in production a-la-MAR (Marshall-Arrow-Romer) can yield concentration of similar amenities in one particular neighborhood. Finally, the presence of migrant population with a particular background can create ethnic amenities associated with that group. In what combinations and proportions these elements are organized throughout the internal geography of a city is a question that has not received much attention by researchers. This is partially because previous studies on the topic usually consider cities in the aggregate and do not recognize the spatial dimension inherent within each urban area. Another apparent reason is the fact that the value of consumer amenities is often measured indirectly through for example land prices (see for example hedonic price models (Rosen, 1974).

We propose to use revealed preference data on actual consumption patterns to explore compositional aspects of urban amenities and identify the main forces and factors behind them. For many cities revealed preference data is not readily available from conventional surveys, census or registry data sources. This might be one of the underlying causes for the absence in the literature of the use of revealed preference for consumption preferences. In this paper, we use a novel dataset from the online location-based service Foursquare that allows us to capture directly preference for certain types of consumer amenities and also provides a wide inventory of other urban amenities.

Foursquare is an application that allows users to checkin at any location and share that location with their online friends. Typically, these locations are public places. Additionally, users can grade and comment on locations which makes the service convenient as a customer-evaluated comparison tool for amenities like restaurants, bars and museums.\footnote{These types of customer-evaluated comparison tools are widespread in many markets like hotels (booking.com), restaurants (tripadvisor.com).} We use the number of checkin’s as an indicator of preferences for a particular amenity in a particular local area. We interpret each checkin as a statement of a positive experience, as online location-based services
and social media are in general biased towards positive experiences, i.e. users show the nice experiences more often than the bad experiences.

The number of checkins from Foursquare are collected over different consumer goods, but we focus on local differentiable non-tradable consumer goods, in particular on restaurants. Unlike for example shops, where you can buy products and consume them at any other desirable location, restaurants closely mirror local demand (see Waldfogel, 2008) and preferences for local characteristics and spatial arrangements of consumer goods. The number of checkins and the spatial distribution of restaurants provide a good indication of the local area compositional effect of popular consumer locations.

In addition to very closely resembling local demand, restaurants are easily differentiable based on the ethnicity of their cuisine. In this paper, we use ethnic restaurant categories to measure local horizontal product differentiation; this is then used to measure the area product compositional effects of popular consumer locations. Population diversity is also relatively easy to identify in a city and can be another source of variety in popular consumer locations\(^2\). Both are also hypothesized to be related and possibly reinforce each other if population diversity causes and enhances product diversity. In recent work by Mazzolari and Neumark (2012), this link is also made as the presence of immigrant populations increases product diversity in terms of ethnic restaurants\(^3\). Restaurant differentiation is used more often as a measure of local product differentiation (see for example Waldfogel, 2008; Mazzolari and Neumark, 2012; Schiff, 2013) but research focusing on the consumer utility from product heterogeneity is however scarce. Waldfogel (2008) finds that, for locally produced and consumed products, the population composition and the local product range are closely related if the population is homogeneous in terms of preferences in these goods.

Although Foursquare provides unique lenses to study consumption, product heterogeneity and the local composition of consumer amenities, the data are not without drawbacks. In particular, two of them appear as the most relevant. First, there is no clear indication of the consumer population represented by the data. There is no detailed socio-economic or demographic information available about the Foursquare users. However we are positive the data contains an inherent bias. The segment of the population who uses this kind of applications is probably overrepresented by young and highly skilled people. The assumption that can be made about these individuals is that they apparently care about letting the world (or their online world) know where they are and what they are doing. There might be thus an advantage in using this group over an unbiased sample in that they frequent restaurants and bars more often, i.e. consume relatively more non-tradeable consumer goods, and hence have more experience that is reflected in their choices. If it is true they derive higher-than-average utility from being in popular places and explore the consumer landscape in a city, this group of people might be of much interest for the ambitions of this study (e.g. implications for urban planning and policy decision). On the other hand, we find the data are not very biased as far as the geographical coverage is concerned, as we show in Section 2.

The second drawback of the data is that no inferences can be made about the causality between the number of checkins and the local spatial arrangements of consumer amenities. This is a common problem in this type of research where past agglomeration and urban patterns interfere with our current observation of spatial structures. However we do believe that exploring the correlations between checkins and consumer good compositions will add to our understanding

---

\(^2\)One can also focus on vertical product differentiation if data on quality of goods is available. Additionally, population differentiation can be measured along the lines of income or educational variety. However, neither of these indicators is available for us.

\(^3\)In accordance with these results, we also find that the presence of immigrants from a specific ethnicity increases the probability of having a restaurant from the same ethnicity. Results are available upon request.
of what makes an attractive consumer city. In various regression specifications we can control for unobserved heterogeneity and omitted variable bias by including neighborhood, local area and restaurant category effects. This greatly enhances our believe that the correlations we find are consistent and robust.

Using this approach, we are able to confirm some established results in the literature about the relation between local product variety as given by Glaeser et al. (2001), but also Fujita et al. (1999) who introduced the concept of consumers’ ‘love of variety’. We provide evidence on the relevance of the spatial structure of amenities within a city for its perceived attractiveness, and contribute some first insights into the role that restaurant diversity and “sector” concentration play in popularity. We find that checkins are higher in areas with horizontal product differentiation and product concentration, indicating that clustering of demand for diversified goods positively relates to the attractiveness of these goods. Although population diversity in terms of ethnicity increases the probability of having restaurant diversity, we only find a relation between the popularity of restaurants and the local population for some ethnic restaurants.

The remainder of the paper is structured as follows. Section 2 we discuss the data we used in detail and relate the data from Foursquare to the data from conventional sources and point out that the data is very comparable and what causes the data from Foursquare to be ’off’. Section 4 gives the regression results that are specified for checkins at the area level, and checkins at the individual restaurant venue level. Section 5 concludes.

2 Data

2.1 Location-based services

One of the main contributions of this study relates to the way we are able to measure revealed preferences for consumers of urban amenities. This is possible thanks to a unique dataset collected from the location-based service (LBS) Foursquare. Before delving into more detail about the collection process, the coverage of the data and the indices derived from them, let us explain what LBSs are and why Foursquare might be the most desirable one for this study.

During the last few decades, the world has witnessed an explosion in computing power that has made possible, among other things, to put a machine that is more powerful than any other one available twenty years ago in the pocket of a non-experienced user. In parallel, location technology such as the global positioning system (GPS) has also undergone dramatic improvements and sharp drops in cost, enabling it to reach the consumer mass. The combination of these two trends is producing a vast amount of geo-referenced data, to an extent that was unimaginable only a few years ago, presenting many opportunities for research in the social and urban realms (Arribas-Bel, 2014). Powered by these technological advances, at the intersection of the Web 2.0 (DiNucci, 1999) and the volunteered geographic information (VGI, Goodchild, 2007) revolution, is the phenomenon called “location-based services”. These are online applications that allow users to broadcast their location in real-time in what has come to be known as checkin’s. These forms of metadata are generated on the user side with location-aware mobile devices (e.g. smartphone or tablet), sent to the central servers of the company through an app and, there, stored in a master database. On the other end, LBSs typically provide application programming interfaces (APIs) that allow third-party developers to query parts of the master database. If the service is popular, this central repository grows extremely large, containing many places such as amenities (e.g. bars, restaurants, parks, theaters) or elements of public (e.g. train stations, bus stops, etc.) and even private infrastructure (e.g. apartments); everywhere subject to be checked in eventually is. These databases then effectively become a reliable digital
representation of some aspects of the physical world. Even more, because not only items from the world but also traces of human behavior (in the form of checkin’s) are stored, they offer unprecedented opportunities to study many of the questions the social sciences are interested in.

There are good reasons to believe LBSs are reaching the popularity needed to become a useful source of information. Its degree of pervasiveness prompted the Federal Communications Commission to issue a recent report (Bureau of Wireless Telecommunications, FCC, 2012). More recently, a survey by the Pew Research Center on the use of these services in the US found that about 75% of adult smartphone owners use some form of LBS; this amounts to 45% of all adults in that country (Pew Research Center, 2013). More specifically, 12% of smartphone users check in using one of this services, which translates into the 7% of all adults. In terms of the demographics of these users, the same report found very few differences in the adoption of this practice among several population groups (only Hispanic had a statistically significant higher use rate), and no statistical difference by gender or educational attainment. Within Academia, LBSs have also been able to gather sympathies and support. Many references have argued that Science is at the dawn of a new era due to the increasing availability of new sometimes geo-referenced datasets (e.g. Lazer et al., 2009, King, 2011). More specifically on the spatial domain, Miller (2010), reviews the rise of new geo-referenced sources of data, with LBSs being among them, in the context of regional science and discusses how these can be mined and analyzed to extract new knowledge. The data themselves have also started to make appearances in scientific applications. As an example of use of LBSs data in the academic domain, Cranshaw et al. (2012) uses checkin’s to re-draw neighborhood boundaries in several American cities, while Cheng et al. (2011) uses similar digital traces to deduce global patterns of human behavior across space.

One of the leaders in the LBS industry is Foursquare. It was created in March 2009 and, by the end of 2013, there were more than 45 million users who had checked in more than 5 billion times (Foursquare, 2013). Its core mechanism is very similar to what has been described above: users sign into an online site that allows them to post their location from an app that transmits the data into a central database that third-party developers can partially query through an API. This is the instrument we use to collect the dataset used in this analysis. In one of its variants, it allows to obtain places (or venues, as they are called) in the surroundings of a specific location. The collection proceeded as follows: during the month of June of 2013, we queried the Foursquare database from a grid of points equally spaced at 50 meters from each other in the developed areas of the Netherlands, and at 500 meters for the rest of the country; this returned the data available from every venue in the vicinity of the pair of coordinates. Figure 1 displays an example of the original locations queried as well as those of the venues we obtained from Foursquare using the famous Museum Square in Amsterdam. After removing duplicate observations, we were left with around 800,000 unique places for which we have access to their name, location, category in the Foursquare classification, how long they have been on Foursquare for, the total count of checkin’s and the number of unique users who had checked in at the venue. In order to obtain a more consistent and complete set of checkin’s count, we reran the query on the specific venues (not on the point grid) in August 2013, and that is what we use throughout the analysis. We adopt the category classification provided by Foursquare to focus only on restaurants and further disentangle which kind of restaurant is each, aggregating the most detailed code by World region. Based on the official one, we develop a more general classification of the restaurants into Western (including Dutch) and non-Western. Table 1 shows the final categories and classifications we use as well as the number of restaurants in each of them, and Appendix ?? contains a detailed relation of Foursquare’s and our own-defined restaurant classes.
NOTE: Black dots correspond with locations in the grid used to query the Foursquare database. Red dots represent Fourquare venues obtained as a response to the queries. Background data come from OpenStreetMap and are available under a CC-By-SA license.

Figure 1: Foursquare data collection. Amsterdam’s Museum Square.

<table>
<thead>
<tr>
<th>Region</th>
<th>Count</th>
<th>Pct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>8,672</td>
<td>16.25</td>
</tr>
<tr>
<td>Eastern Europe</td>
<td>150</td>
<td>0.28</td>
</tr>
<tr>
<td>Middle East</td>
<td>1,585</td>
<td>2.97</td>
</tr>
<tr>
<td>Africa</td>
<td>381</td>
<td>0.71</td>
</tr>
<tr>
<td>North America</td>
<td>10,608</td>
<td>19.88</td>
</tr>
<tr>
<td>Central America</td>
<td>448</td>
<td>0.84</td>
</tr>
<tr>
<td>South America</td>
<td>560</td>
<td>1.05</td>
</tr>
<tr>
<td>Central Asia</td>
<td>341</td>
<td>0.64</td>
</tr>
<tr>
<td>Asia</td>
<td>1,248</td>
<td>2.34</td>
</tr>
<tr>
<td>East Asia</td>
<td>3,770</td>
<td>7.06</td>
</tr>
<tr>
<td>Pacific</td>
<td>31</td>
<td>0.06</td>
</tr>
<tr>
<td>Other</td>
<td>25,575</td>
<td>47.92</td>
</tr>
<tr>
<td>Total</td>
<td>53,369</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 1: Categories used and proportion of venues
2.2 Coverage and bias of Foursquare dataset

Much of the academic discussion and skepticism on LBSs is centered around their biases and coverage (or lack thereof) in representing the underlying general population. These are likely to be present in our Foursquare dataset and, as mentioned in the introduction, their particular direction and subset of the population they potentially over-represent are of high relevance for the purpose of this paper. Despite the interest on this issue, the literature aiming to quantify it is, with exceptions (e.g. Hecht and Stephens, 2014), very scarce. In this section, we benchmark the collected data against two official sources to obtain a representation of where the mismatches can potentially be more pronounced and hence help in a better understanding of Foursquare data.

Finding the right dataset to compare with is necessarily an imperfect endeavor. In fact, the very reason why we believe alternative data is a good research resource makes it a complicated one: there are no official sources to directly measure amenity popularity. This means we are not going to be able to compare checkin activity. However, a tightly connected indicator is the mere presence of venues in the Foursquare database: locations so unpopular so as to not attract a single checkin will not be recorded. Since it is possible to obtain (somewhat crude) restaurant counts from official sources for the entire Netherlands, we compare the distribution of venues in the country, as portrayed by Foursquare, with that from two main sources: building data from the cadaster and data from Statistics Netherlands. We employ different techniques that can best exploit the information provided by these sources. The final picture obtained is rather consistent.

The national cadaster (BAG, for its initials in Dutch) provides access to the location and a few attributes of every single building, and the units that compose it, in the country. One of these characteristics provided is a rough indicator of the function they fulfill. We extract building units devoted to social gathering. Although arguably imperfect, this is the closest match to our restaurant venues. Since the spatial location of these units is given at the point coordinate level, the best way to take advantage of this resolution is a method that does not require any aggregation (e.g. unit count at the neighborhood level), such as kernel density estimation (KDE, Rushton and Tiwari, 2009). This is a standard technique in the spatial analysis of points that essentially computes a probability surface for the location of points. In this context, KDE is a tool to help us compare the spatial distribution of the two sets of points.

Figure 2 (a) and (c) display in a comparable scale the individual distributions of Foursquare and the BAG, respectively. Both maps capture the same general spatial arrangement. The Randstad area to the West appears clearly more populated, while the North of the country is rather empty. However, it also becomes clear both maps are not exactly equal. Upon visual inspection, it is possible to tell the area around Amsterdam is darker in the Foursquare map, while some parts of the north have a higher probability when one uses BAG data. To obtain a more clear comparison, Figure 2 (b) represents the difference between the two maps: BAG estimated probabilities are subtracted from Foursquare ones and plotted in a scale from -1 to 1, effectively providing an indicator of Foursquare over-representation. Specific misalignments become thus clear. The single area with a larger gap between both sources is clearly that around Amsterdam. In this region, we find a notably higher density of Foursquare venues as compared to the underlying BAG benchmark. Differences in the rest of the country are much milder, with the center slightly leaning towards over-representation and the opposite holding true for the North and the top South.

The second comparison uses data from Statistics Netherlands. Statistics Netherlands provides accessibility measures for several types of urban amenities at the neighborhood level, albeit in

---

4 Those under the category bijeenkomstfunctie.
a slightly different way than simple counts: for each area, it is possible to obtain the average number of locations within 3 kilometers by road of all residents in an area. We use this neighborhood index of accessibility to restaurants and set up a sensible comparison with our dataset, aggregated at the same spatial level. Using neighborhood counts of Foursquare venues, we estimate the following regression:

\[
\text{cbs}_i = \alpha + \beta 4\text{sq}_i + \gamma \sum_j w_{ij} 4\text{sq}_j + u_i \tag{1}
\]

where \(\text{cbs}_i\) represents the Statistics Netherlands accessibility measure for neighborhood \(i\), \(4\text{sq}_i\) the number of Foursquare restaurants in the same neighborhood, and \(w_{ij}\) is the \(ij\)-th element of a matrix of spatial weights where \(w_{ij} = 1\) if the centroid of neighborhood \(j\) is closer than 3 kilometer to that of \(i\), zero otherwise and \(w_{ii} = 0\). In other words, this equation is predicting the Statistics Netherlands accessibility index with the combination of Foursquare locations in a given neighborhood, plus those in “roughly” a 3 kilometer buffer.

The upper panel in Table 2 displays the estimates of Equation 1. The coefficients are positive and highly significant. They are also always smaller than one pointing to, on average, a larger number of venues in the Foursquare dataset than in that data from Statistics Netherlands. The critical aspect that we are interested in for this paper is the proportion of variation in the Statistics Netherlands’ data that we are able to explain with Foursquare. A close match between the two will point to a good coverage of the Foursquare data, at least as measured by the official statistics to which we have access. Even bearing in mind the mismatch produced by the differences in the exact definition of the variables, Foursquare data are able to explain more than 90% of the variation in the Statistics Netherlands variable for restaurants.

By examining the error of the model, it is possible to obtain a picture of the bias in an alternative but complementary fashion to the one in Figure 2 (b) using KDE. The mismatch can be due to the methodological differences in calculating the variables outlined above, or to a true lack of proper coverage in the Foursquare dataset. Equally, poor alignment of the two variables may very well vary over space. To assess this situation, we explore the residuals of Equation 1. The bottom panel of Table 2 shows the estimates of regressing \(|u_i|\) on population,

---

5In particular the Statistics Netherlands variable employed is AV3_RESTAU.

6This is indeed less accurate than the Statistics Netherlands measure because we take the geographical centroid, without using any population weighting scheme and either include the entire neighborhood, or discard it, while Statistics Netherlands is effectively including only that part of adjacent neighborhoods exactly within 3 kilometers.
Upper panel shows estimates from Eq. 1; the absolute value of its residuals is used as dependent variable in the bottom panel. Explanatory variables in the bottom regressions are rescaled to per 10,000 units to obtain more readable coefficients.

Table 2: Comparison CBS-Foursquare

<table>
<thead>
<tr>
<th></th>
<th>Restaurants</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Constant</td>
<td>-4.29***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4sq venues</td>
<td>0.54***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>W 4sq venues</td>
<td>0.22***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Constant</td>
<td>6.46***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Population</td>
<td>19.86***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Area</td>
<td>-21.58***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Easting</td>
<td>-0.35***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Northing</td>
<td>0.14***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>11,151</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Comparison CBS-Foursquare

area and coordinates of the neighborhoods, characterizing thus the absolute degree of error in the prediction, irrespective of its sign. Results are highly significant, although explanatory power is substantially lower than in the previous model. Denser and Western areas, such as the dark blue Randstad of the KDE map, tend to see higher disparity with the predicted values; equally, the more north a neighborhood is, the more error in our model.

Although the main component of the analysis is the Foursquare dataset, additional data for Statistics Netherlands is used to include the more traditional socio-economic measures at the neighborhood level. Neighborhoods are the most spatially detailed unit for which this kind of data are available publicly, providing us with the best possible match between Foursquare venues and characteristics of their surroundings. We use total population, average housing values, tenure type, population density and migrant shares. These include a more general Western/non-Western breakdown as well as more detailed percentages by its country of origin (Moroccan, Turkish, Caribbean). In order to be as comparable as possible with the Foursquare dataset, we merge Western and Dutch population and treat it as a single entity.

3 Empirical strategy

Urban diversity is a fairly broad term that loosely refers to the presence of mixed heterogeneity associated with (some) cities. We consider two of its main forms, namely residential and restaurant diversity, and translate them into specific ways to measure them so it becomes feasible to study its potential associations with popularity. The first one relates to the compositional structure of the population who lives in a neighborhood, while the second type focuses on the characteristics of restaurants located in a given part of the city. These variants of diversity

---

aim to capture two of the main factors we hypothesize influence the popularity of an area: the range of available alternatives (restaurants) and its “ethnic profile” (residents). We use the fractionalization index to calculate diversity (Mauro, 1995), whose expression is:

\[
div_i = 1 - \sum_{g=1}^{M} (p_{ig})^2
\]  

where \(div_i\) is the diversity index in area \(i\), \(M\) is the total number of different groups or categories and \(p_{ig}\) is the share of the population from group \(g\) in area \(i\). The groups we use to categorize residents and restaurants are based on nationality and the ethnic categories developed in Table 1, respectively. Intuitively, \(div_i\) represents the probability that two observations randomly selected from area \(i\) belong to different groups. It is bounded \(0 < div_i < 1\), so the closer to 1, the more diverse an area.

Together with diversity, we are also interested in exploring the effect of concentration. We hypothesize that, similarly to how a diverse surrounding can exert a popularity premium, co-location with restaurants of similar characteristics can either give rise or signal an agglomeration of expertise and quality that may also be reflected in bigger traction. The measure of concentration \((C_i)\) for restaurant \(i\) is defined as the proportion of all the restaurants in its surroundings that belong to its same group \(g_i\):

\[
C_i = \frac{\sum_{j=1}^{N} w_{ij} \times K(g_i)_{j}}{\sum_{j=1}^{N} w_{ij}}
\]

where \(w_{ij}\) is the \(ij\)-th element of a spatial weights matrix that assigns 1 if \(i\) and \(j\) are neighbors (including \(i = j\), contrary to common practice in the use of these matrices) and 0 otherwise, and \(K\) is a kernel function that returns 1 if restaurant \(j\) is in the same group \(g_i\) and 0 otherwise. Only the satisfaction of both constraints, spatial and ethnic, results in restaurant \(j\) increasing the local concentration index \(C_i\) for location \(i\).

Considering both diversity and concentration allows us to bring into the discussion two of the most documented and studied types of economies of agglomeration from the urban economics literature: urbanization externalities, named after Jacobs (1961); and the Marshall-Arrow-Romer (MAR) type, also termed as localization externalities. In this context, a diversity bonus could be interpreted, as Jacobs argued, in terms of the benefit of bringing closely together different industries (types of restaurants in this case). Equally, a concentration effect might lend support for the localization argument, in which it is only industries in the same sector that benefit from being close to each other.

We begin our analysis by considering popularity at the neighborhood level. We are concerned with the main determinants and, in particular at this stage, with the role of diversity in determining how attractive an entire area within a city is. We set up the following equation:

\[
\log(ch_a) = \alpha B_a + \gamma F_a + \mu_m + u_a
\]

where \(ch_a\) is the volume of total checkins or total checkins into restaurant venues aggregated at area \(a\), as a measure of popularity; \(B_a\) is a set of controls relating to the neighborhood, such as population and population density, ownership and ethnic structure and average housing value; \(F_a\) are counts of venues in different categories appearing in our Foursquare dataset.
aggregated at the same area level; $\mu_m$ are municipality fixed effects; and $u_i$ is the error term.

The descriptive statistics of the variables included in the regression are given in Table 3. These baseline regressions allow us to explore popularity of entire areas at an aggregate level and thus represent a good first approach to get a feel for the main drivers of popularity.

Table 3: Summary statistics total checkins neighborhoods and restaurants

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Foursquare checkins</td>
<td>5516.876</td>
<td>21928.25</td>
<td>1</td>
<td>874555</td>
<td>6809</td>
</tr>
<tr>
<td>Total Food checkins</td>
<td>726.522</td>
<td>3759.464</td>
<td>1</td>
<td>134944</td>
<td>6809</td>
</tr>
<tr>
<td>$B_P \times 10,000$</td>
<td>0.209</td>
<td>0.235</td>
<td>0</td>
<td>2.715</td>
<td>6809</td>
</tr>
<tr>
<td>$B_P \times 10,000$</td>
<td>0.003</td>
<td>0.003</td>
<td>0</td>
<td>0.025</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>0.229</td>
<td>0.217</td>
<td>0</td>
<td>1</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>0.581</td>
<td>0.251</td>
<td>0</td>
<td>1</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>2.868</td>
<td>1.247</td>
<td>0.78</td>
<td>19.41</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>0.015</td>
<td>0.032</td>
<td>0</td>
<td>0.45</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>0.012</td>
<td>0.029</td>
<td>0</td>
<td>0.5</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>0.015</td>
<td>0.036</td>
<td>0</td>
<td>0.38</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>0.996</td>
<td>0.014</td>
<td>0.743</td>
<td>1</td>
<td>6809</td>
</tr>
<tr>
<td>$B_P \times 100,000$</td>
<td>0.299</td>
<td>0.309</td>
<td>0</td>
<td>1</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>3.361</td>
<td>7.767</td>
<td>0</td>
<td>164</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>7.813</td>
<td>18.338</td>
<td>1</td>
<td>557</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>3.563</td>
<td>8.778</td>
<td>0</td>
<td>233</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>7.233</td>
<td>8.683</td>
<td>0</td>
<td>113</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>15.922</td>
<td>33.115</td>
<td>0</td>
<td>727</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>6.948</td>
<td>33.631</td>
<td>0</td>
<td>2560</td>
<td>6809</td>
</tr>
<tr>
<td>$B_F \times 100,000$</td>
<td>66.889</td>
<td>117.644</td>
<td>1</td>
<td>3547</td>
<td>6809</td>
</tr>
</tbody>
</table>

In order to uncover detail buried in aggregation, we perform a second analysis where we consider individual restaurants. In this case, we benefit from the availability of high spatial resolution and exploit variation in the immediate surroundings of the restaurant. We draw a 500m. buffer around each venue under “Food”, the basis of these micro-regressions, and obtain counts of other Foursquare venues as well as of the type of restaurants in the vicinity. The equation to estimate becomes more complete and allows us to control for a larger part of previously unobserved restaurant heterogeneity:

$$\log(ch_i) = \alpha F_i + \beta S_i + \gamma B_{ai} + (\mu_m) + (\eta F_i \times B_i) + u_i$$

where $ch_i$ is the total volume of checkins in restaurant venue $i$; $S_i$ is the count of other Foursquare venues in the 500m. surrounding $i$, including the diversity and concentration measures; $B_i$ is a set of variables relating to the neighborhood $a$ in which $i$ is located, or a set of area fixed-effects; in cases where we do not consider area fixed effects, municipality fixed effects $\mu_m$ are estimated. Finally, some regressions incorporate the term $\eta F_i \times B_i$ that interacts the restaurant category with the proportion of population with that ethnic background in the neighborhood. This can be interpreted as a different form of concentration that also appears interesting to explore. We hypothesize that restaurants of a specific cultural background will signal or benefit\(^8\) from being in areas with a higher population with that specific background.

\(^8\)As already mentioned, at this stage, we cannot distinguish a causal link from sorting or a “survival of the fittest” type of process. Hence, a positive and significant coefficient in this term could equally point to an effect of locating in neighborhoods with higher populations of the same cultural background, or simply signaling that
(e.g. Turkish restaurants will be more popular in neighborhoods with more Turkish resident population) because of the presumably more demanding and experienced set of customers they directly cater. This idea has been proposed by Waldfogel (2008), among others, who argues that particular varieties of a product will emerge only if there is a large number of nearby consumers favoring the product.

4 Results

4.1 Popular areas

In the analysis of factors that constitute a popular area composition we distinguish between the analysis at the neighborhood level, and the analysis over the aggregated restaurant checkins at the neighborhood level. We start with the former as is given in Equation 4. From regression 1 and 2 in Table 4 we can get the broad picture of what aspects of neighborhoods are related to a higher number of overall checkins, i.e. what areas are popular among Foursquare users. A municipality fixed effect is included so unobserved level differences between cities are accounted for. Before turning to our main variables of interest, the product differentiation and population diversity, we discuss the results of the covariates.

Although the covariates serve as control variables, their association with the number of checkins brings about additional information on what neighborhood characteristics make popular areas. Including the total number of Foursquare venues as well as the size of the neighborhood measured by the population, we find that, unsurprisingly, larger areas have higher number of checkins. However, denser populated areas do not have more checkins. An explanation for this result can be that residential areas in general do not always coincide with dense consumer amenity areas. The economic characteristics of a neighborhood all show a negative association with the total number of checkins. Higher income neighborhoods thus have less checkins.

Turning to the Foursquare categories, neighborhoods with a higher number of venues in the “shop and service” and “outdoors and recreation” categories have higher number of checkins all other things equal, while the number of “food” and “arts and entertainment” venues and “travel spots” have a negative correlation with the number of checkins. The size of the estimated coefficients of the different categories differs substantially. An extra “shop or service” spot is related to an increase in checkins of 0.5%, an additional “recreational” spot is related to a 4.3% increase in checkins (all other things equal). An additional restaurant or “Food” venue also has an higher relation with the number of checkins, namely a decrease of 2.9% and an additional venue in “arts and entertainment” is related to a decrease in checkins of 1.7%. It seems that although the largest Foursquare category in our dataset is “Shops & Services”, the number of shops has much less of an impact on the number of checkins than the other categories.

At the aggregated neighborhood level, checkins are higher if product diversity measured as restaurant diversity is larger and the effect diminishes if the diversity of restaurants gets higher suggesting that there might be some optimum number of diverse restaurants for attractive areas (in regression 2 in Table 4). An increase in restaurant diversity of 0.1 (from for example 10% to 20%), is associated with an increase in checkins of 17% all things equal. The diversity and composition of the neighborhood population is not related to popularity of neighborhoods. It thus seems that at a higher aggregate level, the overall popularity of a neighborhood is not related to the diverse composition of the local population, but only to the diversiy in products. Although we know from research and our own analysis of the data that the probability of having those located there hold unobserved characteristics that make them more popular.
restaurant diversity is related to the local population diversity, we do not find an additional
popularity premium for neighborhoods from a diverse population.

These results also hold when we turn to analyzing the total number of checkins to restaurants
in a neighborhood. Popular Foursquare restaurants are located in neighborhoods that have a
diversified supply of restaurants, even when we control for the number of restaurants. As we
control for the total number of venues, the positive and significant coefficient of the number and
variety of restaurants indicates that both clustering and product differentiation are positively
related to the popularity of restaurants in an neighborhood. An additional restaurant is associated
with an increase in checkins into restaurants of 2%, while an increase in restaurant diversity of
0.1 is associated with 35% more checkins into restaurants.

Table 4: Full area regressions

<table>
<thead>
<tr>
<th></th>
<th>(1) checkins_total</th>
<th>(2) checkins_total</th>
<th>(3) checkins_restaurants</th>
<th>(4) checkins_restaurants</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_population (× 10,000)</td>
<td>0.475***</td>
<td>0.490***</td>
<td>-0.185</td>
<td>-0.164</td>
</tr>
<tr>
<td>B_Pop_density</td>
<td>-37.667***</td>
<td>-36.965***</td>
<td>15.082</td>
<td>16.030</td>
</tr>
<tr>
<td>B_P_socialrent</td>
<td>-0.293**</td>
<td>-0.298**</td>
<td>-0.275**</td>
<td>-0.281**</td>
</tr>
<tr>
<td>B_P_owner_occupied</td>
<td>-0.604***</td>
<td>-0.612***</td>
<td>-0.506***</td>
<td>-0.518***</td>
</tr>
<tr>
<td>B_taxvalue (× €100,000)</td>
<td>-0.048**</td>
<td>-0.048**</td>
<td>-0.005</td>
<td>-0.005</td>
</tr>
<tr>
<td>B_P_Caribbeans</td>
<td>0.832</td>
<td>0.860</td>
<td>-1.296</td>
<td>-1.258</td>
</tr>
<tr>
<td>B_P_Moroccans</td>
<td>-0.572</td>
<td>-0.550</td>
<td>-2.613*</td>
<td>-2.578*</td>
</tr>
<tr>
<td>B_P_Turks</td>
<td>-0.582</td>
<td>-0.575</td>
<td>-1.190</td>
<td>-1.180</td>
</tr>
<tr>
<td>B_Fractionalization</td>
<td>0.873</td>
<td>0.933</td>
<td>-4.216</td>
<td>-4.134</td>
</tr>
<tr>
<td>F_P_FastFood</td>
<td>-0.014</td>
<td>-0.019</td>
<td>0.017</td>
<td>0.009</td>
</tr>
<tr>
<td>F_Arts &amp; Entertainment</td>
<td>-0.017***</td>
<td>-0.017***</td>
<td>-0.013*</td>
<td>-0.013*</td>
</tr>
<tr>
<td>F_Food</td>
<td>-0.029***</td>
<td>-0.029***</td>
<td>0.021***</td>
<td>0.022***</td>
</tr>
<tr>
<td>F_Nightlife Spot</td>
<td>-0.001</td>
<td>-0.001</td>
<td>-0.030***</td>
<td>-0.036***</td>
</tr>
<tr>
<td>F_Outdoors &amp; Recreation</td>
<td>0.032***</td>
<td>0.032***</td>
<td>0.026***</td>
<td>0.026***</td>
</tr>
<tr>
<td>F_Shop &amp; Service</td>
<td>0.005***</td>
<td>0.005***</td>
<td>0.005**</td>
<td>0.005**</td>
</tr>
<tr>
<td>F_Travel &amp; Transport</td>
<td>-0.009***</td>
<td>-0.008***</td>
<td>-0.006***</td>
<td>-0.006***</td>
</tr>
<tr>
<td>F_total</td>
<td>0.009***</td>
<td>0.009***</td>
<td>0.004***</td>
<td>0.004***</td>
</tr>
<tr>
<td>F_fractionalization</td>
<td>1.731***</td>
<td>2.153***</td>
<td>3.459***</td>
<td>4.029***</td>
</tr>
<tr>
<td>F_fractionalization²</td>
<td>-0.670**</td>
<td>-0.903*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>6.217**</td>
<td>6.142**</td>
<td>7.544*</td>
<td>7.443*</td>
</tr>
<tr>
<td>B_gem</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Adjusted $R^2$</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>0.633</td>
<td>6809</td>
</tr>
<tr>
<td>(2)</td>
<td>0.633</td>
<td>6809</td>
</tr>
<tr>
<td>(3)</td>
<td>0.529</td>
<td>6809</td>
</tr>
<tr>
<td>(4)</td>
<td>0.529</td>
<td>6809</td>
</tr>
</tbody>
</table>

Standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

One additional explanation for the attractiveness of restaurants and restaurant locations
can be the quality or pricing of the restaurant. Although we cannot control for this vertical
product differentiation, including an indicator for the share of the “Food” venues that is fastfood
(self-serving) and therefore lower-end priced or not will partly deal with this. In the regressions
in Table 4, this has no significant relation with the number of checkins of the neighborhood nor
with the number of checkins into restaurants.

An aggregated checkin analysis at the neighborhood level does not allow to control for many
unobserved characteristics of neighborhoods, like other amenities or aesthetic views or buildings,
and of restaurants that could drive the above found results. We therefore perform the analysis
on individual restaurants and their direct surroundings. The focus of the analysis then shifts
from the components of overall neighborhood popularity to the neighborhood components that
relate to the popularity of a restaurant.
4.2 Popular venues

The regression results for the individual restaurant checkins are given in Table 5. The first two columns include municipality fixed effects while the third one includes a neighborhood fixed effect that is further explained in the fourth column. The first regression additionally includes restaurant cuisine-region effects, i.e. fixed effects for the region from which the cuisine originates, controlling for the unobserved variation that can exist between restaurants of different origins. If, for example, European restaurants are more higher-end than the Asian ones, or get more checkins because they are more likely to be on Foursquare, this is captured by the region category fixed effect. In the second regression we do not include a region category fixed effect, but separately look at the ethnic origin of the cuisine of restaurants in relation to local presence of the population from the same ethnic origin.

In general, the covariate results are comparable to the neighborhood regressions, with population density, the share of owner occupied dwellings and social rent having a negative relation with the number of checkins into a restaurants. Unlike in the neighborhood specification\footnote{The difference in results can be due to the different unit of analysis and neighborhood popularity is unrelated to overall neighborhood economic characteristics while restaurant popularity is not. However, the results of the neighborhood analysis could also be influenced by unobserved neighborhood heterogeneity within municipalities or omitted variable bias.}, the positive coefficient of the average taxation value of the dwellings surrounding the restaurant suggests that restaurants are more popular in expensive areas. If local housing prices signal local income levels, and local population preferences mirror the local availability of non-tradable consumer goods, this is an indication that more expensive restaurants are more popular. The negative coefficient on the fast food dummy points into the same direction. The local share of Turks and Moroccans has a significant negative correlation with the number of checkins into a restaurant while population diversity has no significant relation.

The results for the concentration and fractionalization indices show that both restaurant diversity as well as the clustering of restaurants of the same cuisine are related to a higher volume of checkins. If the concentration of restaurants from the same ethnic background in the surrounding of a restaurant increases with 0.1, the associated increase in the volume of checkins into that restaurant is almost 25%, all other things equal. An increase of 0.1 in the diversity of the surrounding restaurants is associated with 27% increase in the volume checkins of a restaurant. Restaurants are popular in areas that offer choice among different restaurants, which is an extended result of what is hypothesized by Glaeser et al. (2001) and also by Ottaviano and Peri (2006). However, having closeby competition of restaurants of the same regional kitchen is also related to higher popularity. This points to positive competition and agglomeration externalities in terms of quality or price, but also in terms of production if the population of a certain ethnic background has a (comparative) advantage in supplying food from their own ethnic background. In addition to clustering based on the supply side of the market, the demand side can play a role too. For example, the demand for Turkish restaurants is high in neighborhoods with many Turks. This hypothesis is tested in regression 2 in Table 5.

The results of the second regression in Table 5 closely resemble those of the first one, but allow us to obtain an additional insight into whether restaurants with a specific ethnic background are more popular if they are located in an area with a higher population of that ethnic background. Our data only allows us to distinguish between Turks, Moroccans and Caribbeans. For the last two, the presence of residents with those national backgrounds, has no significant relation with the popularity of Caribbean or Moroccan restaurants. For Turkish restaurants, we find they have higher checkins in areas where population with the same background is more abundant. Turkish restaurants experience a checkin premium if they are located in Turkish neighborhoods. At least for this group, we find suggestive evidence of positive agglomeration, or competition,
Table 5: Full venue regressions

<table>
<thead>
<tr>
<th></th>
<th>0-Restaurants</th>
<th>1-Restaurants</th>
<th>2-Restaurants</th>
<th>3-Buurt FE</th>
</tr>
</thead>
<tbody>
<tr>
<td>B_population (× 10,000)</td>
<td>0.033</td>
<td>0.029</td>
<td></td>
<td>0.013</td>
</tr>
<tr>
<td>B_P_socialrent</td>
<td>-0.003***</td>
<td>-0.003***</td>
<td>-0.007***</td>
<td></td>
</tr>
<tr>
<td>B_P_owner_occupied</td>
<td>-0.005***</td>
<td>-0.005***</td>
<td>-0.008***</td>
<td></td>
</tr>
<tr>
<td>B_taxvalue (× €100,000)</td>
<td>0.045***</td>
<td>0.044***</td>
<td>0.036*</td>
<td></td>
</tr>
<tr>
<td>B_P_Caribbeans</td>
<td>0.002</td>
<td>-0.001</td>
<td>0.005</td>
<td></td>
</tr>
<tr>
<td>B_P_Moroccans</td>
<td>-0.006**</td>
<td>-0.008**</td>
<td>-0.006</td>
<td></td>
</tr>
<tr>
<td>B_P_Turks</td>
<td>-0.007***</td>
<td>-0.01***</td>
<td>0.004</td>
<td></td>
</tr>
<tr>
<td>B_Fractionalization</td>
<td></td>
<td>-0.947</td>
<td>0.813</td>
<td></td>
</tr>
<tr>
<td>F_Caribbeans</td>
<td></td>
<td>-0.442***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F_Moroccans</td>
<td></td>
<td>-0.54***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F_Turks</td>
<td></td>
<td>-0.043</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F_age</td>
<td>0.087***</td>
<td>0.087***</td>
<td>0.087***</td>
<td></td>
</tr>
<tr>
<td>F_FastFood</td>
<td>-0.052***</td>
<td>-0.069***</td>
<td>-0.062***</td>
<td></td>
</tr>
<tr>
<td>S_Arts &amp; Entertainment</td>
<td>-0.002**</td>
<td>-0.002**</td>
<td>-0</td>
<td></td>
</tr>
<tr>
<td>S_Food</td>
<td>-0.003***</td>
<td>-0.003***</td>
<td>-0</td>
<td></td>
</tr>
<tr>
<td>S_Nightlife Spot</td>
<td>0.001</td>
<td>0.001</td>
<td>-0.003*</td>
<td></td>
</tr>
<tr>
<td>S_Outdoors &amp; Recreation</td>
<td>-0</td>
<td>-0</td>
<td>-0.003*</td>
<td></td>
</tr>
<tr>
<td>S_Shop &amp; Service</td>
<td>0.001***</td>
<td>0.001***</td>
<td>0.001*</td>
<td></td>
</tr>
<tr>
<td>S_Travel &amp; Transport</td>
<td>0.001*</td>
<td>0.001*</td>
<td>0.001</td>
<td></td>
</tr>
<tr>
<td>S_Total</td>
<td>0.001**</td>
<td>0.001***</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>S_Concentration</td>
<td>0.249***</td>
<td>0.33***</td>
<td>0.217***</td>
<td></td>
</tr>
<tr>
<td>S_Fractionalization</td>
<td>0.265***</td>
<td>0.276***</td>
<td>0.218***</td>
<td></td>
</tr>
<tr>
<td>F × B_P_Caribbeans</td>
<td>0.004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F × B_P_Moroccans</td>
<td>-0.021</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F × B_P_Turks</td>
<td>0.024***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F_buurt</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>B_gem</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Region Category</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>-</td>
</tr>
<tr>
<td>R²</td>
<td>0.354</td>
<td>0.352</td>
<td>0.335</td>
<td>0.015</td>
</tr>
<tr>
<td>N</td>
<td>43066</td>
<td>43066</td>
<td>46990</td>
<td>4387</td>
</tr>
</tbody>
</table>

Standard errors in parentheses. * p < 0.05, ** p < 0.01, *** p < 0.001

Finally we compare restaurants within the same neighborhood while controlling for unobserved heterogeneity between restaurants of different cuisines, by including a neighborhood and region category fixed effect. The results point to overall robust and consistent findings as the estimated coefficients remain the same. However, the found negative association of the presence of Turks and Moroccans with the popularity of a restaurant is now mainly explained by unobserved characteristics of the neighborhoods as their share is not significant in explaining the neighborhood fixed effects.
5 Conclusion

In this paper we have used a unique dataset of Foursquare checkins to look into the “ingredients” of an attractive consumer city, as well as into its “recipe”, the spatial arrangement of consumer goods within cities by using a unique dataset of Foursquare checkins. We find a robust and consistent association between product diversity of local non-tradeable goods and popularity, both at the area, the aggregated restaurant, and individual restaurant level. This underwrites one of the main hypothesis of attractive consumer areas put forward by Glaeser et al. (2001) and Fujita et al. (1999). In addition to, for example Ottaviano and Peri (2006), who indirectly find the same effect on the city level, our results indicate that the local consumer good composition can have a positive and substantial association with the popularity of that area. Concentration of homogeneous products also positively relates to the popularity of restaurants indicating presence of MAR-agglomeration externalities. The analysis reveals the existence of a subtle relation between product and population heterogeneity in which demand-side effects might impact the popularity of consumer amenities. With this analysis we use the overall relation between product and population diversity as shown by Waldfogel (2008) and show that the popularity of restaurants with a particular ethnic background can depend on the presence of a population from that same ethnic background.

This paper also shows that data derived from location-based services like Foursquare can serve as an informative addition to conventional data on consumer behavior or local consumption patterns. Although the data from these kind of applications is most probably biased, as the users are not randomly selected from the consumer population, the type information obtained from services like Foursquare is valuable for researchers, urban planners and municipalities as revealed consumer data, for example. This is of particular relevance given the increasing importance of local consumer amenities and their role in city development policies.

This research predominantly depends on the horizontal product differentiation of restaurants, which is, as stated in the introduction, “the best we can do” with the available data. However insights into local amenity composition based on quality or prices, and the use of other consumer products and local relative consumer good composition would be valuable extensions for future research. Our analysis of the individual restaurants already deals with the drawback of predetermined administrative areas by composing local areas based on a fixed radius. However, part of the analysis of popular areas should take into account that an area is not necessarily a predefined administrative region but, especially when looking at dynamic phenomena such as popularity, should be endogenously determined. As this is an exercise worth the entire focus of a project, we warrant this for future research.

References


Sharing Services, Proceeding of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM), Barcelona.


Getting out of the pool: shrinkage methods for tall and wide datasets. An example from building energy consumption.

David Hsu

Department of City & Regional Planning
University of Pennsylvania
210 S 34th Street, Philadelphia, PA 19104

Abstract

This paper explores the challenges of tall and wide data, a general class of big data problems in many fields such as genomics and marketing, as well as in urban informatics. It begins by discussing challenges inherent in the structure of high-dimensional or big data, and how the usual goals of analysis, including prediction, explanatory and interpretive power, and parsimony, all are interrelated and compete with one another. Focusing on two particular techniques, clustering and regression, that are increasingly used together in sequence, this paper argues that variable or feature selection is a key aspect of both techniques, and therefore seeks how to specify parsimoniously both regression models for prediction and clustering models for explanatory groupings. Approaches such as shrinkage using the lasso method, and cluster-wise regression using finite mixture models are examined, and applied to a comprehensive, large dataset of building energy consumption in New York City multi-family buildings. It is found that the increasingly common two-stage approach of clustering and then regression overfits the data and ignores statistical uncertainties in the clustering process, while a combination of existing methods results in predictions that are more accurate and clusters that are more stable. This paper concludes with suggestions for future research into sparse cluster-wise regression.

Keywords:
big data, building energy, statistics, feature selection, clustering, regression

Email address: hsuyd@design.upenn.edu (David Hsu)
1 Introduction
2 Analysis goals for big data
3 Related work
4 Theory
  4.1 Considerations in clustering
  4.2 Importance of feature selection
    4.2.1 Clustering for explanation
    4.2.2 Regression for prediction
    4.2.3 Cluster-wise regression
  4.3 Integrated approach
5 Application
  5.1 Data
  5.2 Methods
  5.3 Results & discussion
  5.4 Limitations
6 Conclusions
7 References
1. Introduction

Structured and unstructured data can often be represented as two-dimensional matrices, with \( n \) units or observations, and \( p \) features. Tall data is where \( n \) is large, and wide data is where \( p \) is large. In supervised learning, one of more of the \( p \) features are identified as outcomes of interest, and the others can be either predictors, covariates, or entirely spurious noise. Unsupervised learning approaches are considered to be exploratory analysis methods that do not identify any particular outcome of interest but instead describes relationships among the features.

High-dimensional (or big) data raises a number of challenges for the statistical methods typically used in the social and natural sciences. First, in wide data, there is the supervised learning problem of feature selection: while big data greatly increases the dimension of \( p \), including additional features does not necessarily increase predictive accuracy or insight. Second, regression in high-dimensions when \( p >> n \) can lead to high variance or overfitting of models, which in turn degrades accurate predictions on test data that is either new or from out-of-sample. Third, models containing many features are difficult to interpret, and parsimony is an important value to strive for. Fourth, in tall datasets there is the unsupervised problem of clustering or classification, because large datasets are usually composed of heterogeneous subgroups. Within the overall population, whether it is among individuals, households, or buildings, multiple latent groupings may be distributed nonlinearly, highly correlated with one another, and/or include outliers, which can all lead to highly biased results if standard statistical assumptions are used.

Many of these feature selection and clustering problems are also interrelated. If key features cannot be identified, then it is also difficult if not impossible to identify subpopulations of interest. If subpopulations are not first identified, then predictions on the entire population are often not accurate enough to use within them or else lead to problems in aggregation. All of these challenges force us to reconsider how to balance goals that commonly compete in statistical inference and interpretation, including predictive power, explanatory power, and parsimony.

This paper explores how some of these particular challenges raised by big data should be addressed with integrated statistical approaches. The paper begins by discussing some fundamental goals in analyzing big data, then shows how recent work in the building energy consumption literature that applies common statistical techniques in multiple stages may overfit the data, resulting in incorrect interpretation of statistical errors. It then discusses some of the key strands of theoretical literature that need to be joined into a fully-integrated, principled statistical approach.

An approach is then synthesized from two previous analyses [21, 22, both in submission]
in order to demonstrate how this approach would be applied to a large dataset of observed building energy consumption in New York City multifamily buildings. This is an important environmental and policy analysis issue because buildings consume approximately 40% of total primary energy worldwide, and even more in some cities: for example, in New York, buildings are estimated to constitute 55% of energy use and 75% of greenhouse gas emissions. Interest has grown in statistical models applied to observed energy use in buildings, and has also been aided by mandatory energy benchmarking and audit laws throughout the United States and Europe that are rapidly creating new datasets for analysis and further policymaking.

The paper concludes by discussing how other theoretical methods might be further integrated to analyze the general class of tall and wide datasets in big data and urban informatics problems.

2. Analysis goals for big data

Big data – data that is both taller and wider than previously available – forces us to reconsider fundamentally the goals of quantitative analysis. One could make a reasonable argument that big datasets are in a particular sense almost complete, that is, that they measure everything within certain structured contexts because those contexts could not exist without computer technologies and sensing. For example, Facebook and Amazon may have thousands of observations per user on their websites (via “clickstreams”), and it is hard to argue that any other social networks or shopping portals currently have the same scope. Or, that thousands of measurements per minute from sensors on a jet engine can represent the physical state of the engine. Or, if any national security agency were to be interested in such a thing, then monitoring the location and all of the calls and data transmitted to a smartphone represents a relatively high proportion of what information a person could be receiving other than through face-to-face interaction. These arguments are not so far-fetched or hard to believe.

However, inference and interpretation from data will still remain a fundamental activity for two reasons. First, “complete” big data is still hard to come by, and the number of structured contexts that can be measured to high fidelity and resolution is relatively small (but rapidly growing). Second, even if we have many precise measurements, we will have to interpret data because not all activity of interest occurs in highly structured settings. The quote usually attributed to Einstein [28] comes to mind: “not everything that counts can be counted, and not everything that can be counted counts”. For example, a smart electricity meter, even with a high frequency of communications, may never explain why people consume particular amounts of energy at certain times. Similarly, it is unlikely that
people’s behavior on a social network for dating will fully reflect their attitudes about love and sex, even if those interactions may reveal new truths about how groups of people behave with respect to one another. So we may always need to join and federate other datasets from other units of analysis, like supplementary household behavior surveys and so on, to make sense of what our sensor data is telling us.

Inference and interpretation of big data should therefore have four main goals:

*Appropriateness for high-dimensional data.* The curse of dimensionality manifests itself in many ways. Some of the problems were mentioned in the introduction, including feature selection when $p$ is large, finding subgroups when $n$ is large, and avoiding overfitting when $p >> n$. It is possible to over-specify and over-fit models, by adding too many predictors, so that the analytical power of models becomes saturated: for example, given enough predictors and their linear combinations, a proportion of explained variance ($R^2$) value of 1 is relatively easy to obtain, and given a large enough sample size, all $p$-values become statistically significant and yet practically meaningless. Furthermore, Hastie et al. [16, page 22] note that high-dimensions make it difficult to define any local region, because as the dimension of the data grows, sampling means that every point becomes progressively closer to the outside boundaries of the multi-dimensional boundary sphere. Thus, when carrying out clustering or nearest-neighbor analyses on high-dimensional data, every data point begins to look like its own subgroup.

*Accurate prediction of key metrics or outcomes.* One of the main promises of bigger data is that it will enable us to predict outcomes or effects more accurately. Measurement of prediction error is therefore an important aspect of model validation, and a relatively strong method for comparing the relative efficacy of models is to use cross-validation, where the model is trained on one part of the dataset, and tested on another. Multiple cross-validation will divide the number of observations into folds, say into thirds; train a model on the two-thirds of the data; test on the remaining one-third, and then take the mean-squared error across all folds. Sometimes the dataset will be divided into an additional portion for validation, since implicitly the model is being improved in the testing process as well. Of course, cross-validation only simulates how models will perform in out-of-sample predictions when assuming that out-of-sample data will have the same structure as our original dataset, and that our modeling assumptions remain intact in the future.

*Explanatory and interpretive power.* Outcomes of interest, and the predictors or features used to explain them, should be clearly related to the goal of the analysis. A fundamental strategy in exploratory data analysis is also clustering of subsets within the overall dataset, to explain heterogeneity and difference between parts of the dataset. As Jain [23] writes, “organizing
data into sensible groupings is one of the most fundamental modes of understanding and learning” (page 651). It is desirable that our analyses are also stable, in the sense that if our currently held assumptions continue to hold true, that new data and subsequent analysis will give us similar results, explanation, and interpretation.

**Parsimony.** Finally, and related to the previous three points, our analyses should be parsimonious where possible. While it is very easy to calculate the effects of tens, hundreds, or thousands of variables on a particular outcome, this makes it difficult to interpret and understand. This may also obscure particular decisions made during the analytical process. Parsimonious models are easier to interpret, understand, communicate to others, and maintain. For example, while there may be little or no additional cost to adding variables to an analysis from an existing dataset, these results may not be generalizable to future circumstances unless all of those variables continue to be collected in the future as well. This may lead to considerable ongoing costs in data collection, cleaning, and analysis.

### 3. Related work

As datasets continue to grow rapidly, there has been a great deal of work, particularly in the building literature, that seeks to use clustering for explanatory analysis and then applies subsequent regression models to predict within each group. Clustering is often used as an exploratory technique to improve the results of other methods as well, since applying homogeneous modeling assumptions to a heterogeneous group will usually lead to relatively worse predictions. In addition, many modeling techniques are particularly sensitive to relative outliers, which is really just another way of describing heterogeneity, though the connections between robust statistics and clustering remain relatively unexplored [17]. Therefore it has become relatively common to use clustering as a pre-processing step, in which similar groups are identified in order to model them separately in a subsequent regression analysis [38]. This paper will review just the instances in the building consumption literature in which clustering is used with a subsequent regression analysis, which is still a large subset of a larger set of clustering papers:

**Archetypes.** “Archetype” buildings are sometimes developed in the literature, in order to model them separately based on other knowledge of the general population of buildings, to model an aggregate population measure (such as total energy consumption by buildings), or for particular policy purposes. This is often developed in a ad hoc or non-quantitative way, and then regression models are applied to the sector as a whole. Recent examples include [5, 9, 41].
**Dimensional reduction.** Methods such as principal component analysis (PCA), principal components reduction (PCR) are sometimes used to reduce dimensions of the data, by forming multidimensional scales or eigenvectors that describe the principal sources of variation. With the smaller number of principal components that are linear combinations of other features, sometimes regression is carried out directly, and sometimes clustering is carried out and then regression is applied [recent examples include 13, 8, 39]. Partial least squares (PLS) also includes the outcome itself (the \( y \) vector) in order to construct the principal components. However, Frank and Friedman [11] generally find that ridge regression, which will be discussed below, dominates these methods in prediction error. In addition, these methods do not actually discard or remove any information, since multi-dimensional principal components are simply composed of linear combinations of other predictors. Therefore, this area will be ignored for the rest of this paper.

**Clustering.** Methods such as \( K \)-means and hierarchical clustering are still often used on quantitative data. Yet Jain [23, page 651] writes, “in spite of the fact that \( K \)-means was proposed over 50 years ago and thousands of clustering algorithms have been published since then, \( K \)-means is still widely used. This speaks to the difficulty in designing a general purpose clustering algorithm and the ill-posed problem of clustering”. Jain [23, page 653] further identifies three main uses of data clustering: to understand the underlying structure of the data, to classify items by similarity, and to compress and summarize data through archetypes. Sometimes these are just summarized as clusters of interest [43], and sometimes these same clusters are then modeled using simulation models [30], regression methods [1, 42], or other explanatory techniques [35].

**4. Theory**

From these methods, this paper will use some of the problems with the practice of clustering and modeling separately to motivate the development of a better procedure and method. Many of these problems stem from the way that clustering is used as a pre-processing step. Feature selection in particular plays an important role in achieving prediction, parsimony, and explanatory power. While the unsupervised and supervised problems are often thought of separately, they can also be combined in a principled statistical approach called clusterwise regression.

**4.1. Considerations in clustering**

First, the purposes of clustering should determine the analytical method [23, 38]. Second, there are hundreds of possible clustering algorithms which give very different results.
Third, which features are selected for clustering can result in very different clustering results. Fourth, the natural shape of the data matters: it is well known that applying some clustering algorithms even to homogeneous datasets can still give distinct clusters. Fifth, and building on the previous problems, the process of clustering still contains statistical uncertainty, but very rarely have any subsequent modeling approaches taken this into account, if ever. The following section shows how the issue of feature or variable selection affects both the clustering and prediction stages.

4.2. Importance of feature selection

4.2.1. Clustering for explanation

The clustering literature has extensively addressed the role of initial assumptions in determining clustering results. Key technical concerns have been the number of clusters that one is looking for versus those which may naturally exist in the data; how the data is represented and normalized; the shape and structure of clusters; the elimination of outliers; and the overall clustering objectives. Some but not all of these concerns have been addressed in the very large literature on clustering.

For the particular problem of feature selection, Jain [23] argues that “if the representation (choice of features) is good, the clusters are likely to be compact and isolated and even a simple clustering algorithm such as K-means will find them. Unfortunately, there is no universally good representation: the choice of representation must be guided by the domain knowledge” (page 656). However, this seems to lend credence to the argument by von Luxburg et al. [38, page 75] that “in most uses of clustering, one has a ‘bias’ concerning what one is looking for. This bias affects the type of clusters one tries to construct”.

More technical studies have been done on variable selection. Steinley and Brusco [34] offer a large number of references that recognize that not all variables contribute equally to defining cluster structure, and that the addition of variables that do not define cluster structure can actually degrade the ability of clustering procedures to recover cluster structures. They go on to test a number of clustering algorithms or methods for their sensitivity with regards to variable selection. Gao and Hitchcock [14] detail a number of variable screening and selection procedures used to improve K-means clustering in particular. Kriegel et al. [24] review a number of recent developments in feature selection for clustering.

4.2.2. Regression for prediction

We start out with the standard linear regression problem:

\[ Y = X\beta + \epsilon \]
The typical assumptions are that vector $Y$ is of length $i = 1 \ldots N$, vector $\beta_p$ is of length $p = 1 \ldots p$, the features are a matrix $X$ of $n \times k$ dimension, and the random variable is normally distributed, that is $\epsilon_i \sim N(0, \sigma^2)$. This can be estimated in closed form using ordinary least squares (OLS) regression, in which the sum of the squared error is minimized by the coefficients.

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

(2)

The advantages of OLS are well known, including un-biased coefficients, a closed form solution, and therefore computational simplicity. However, even old reliable OLS requires us to make choices when implemented in high dimensional settings where $p >> n$, particularly with regards to variable selection. Though it is rarely desirable to include all of the available variables for reasons of parsimony, interpretability, statistical significance, or simply space in journal articles, it is not necessarily easy to tell which variables should be included. There is also a tradeoff between practical and statistical significance. When $n$ is very large, and the variables are compared on a standardized basis, it is easy to get significant $p$-values for almost every coefficient even if the coefficients are relatively small. It is generally more desirable to identify the highest-magnitude standardized coefficients, which have the greatest impact on the outcome, than to include many tens if not hundreds of additional coefficients which may represent relatively minor effects on the outcome. In addition, as $p$ increases, the possible number of subset models that include or do not include particular variables increases rapidly ($2^p$). For a dataset with 100 features, this leads to $1.26 \times 10^{30}$ possible subset models. Furthermore, this also increases the number of possible interactions, which increases as the binomial coefficient and which are frequently of interest. Using the same hypothetical dataset, if we were to select the “full” model of 100 features, then the number of possible interactions is $\binom{p}{2}$, or 4,950 additional interactions between the features. This would of course be even larger if we were to examine all of the possible number of interactions within all of the subsets.

Finally, an important problem with OLS is our estimation of unbiased coefficients comes at a price. The bias-variance tradeoff decomposition states that the mean squared error of our predictions $\hat{f}(x_0)$ can be restated as the sum of the variance of the predictions, the bias of the predictions squared, and the variance of the error terms [16]:

$$E \left( y_0 - \hat{f}(x_0) \right)^2 = \text{Var} \left( y_0 - \hat{f}(x_0) \right) + \left\{ \text{Bias} \left( y_0 - \hat{f}(x_0) \right) \right\}^2 + \text{Var}(\epsilon)$$

(3)

Bias falls off rapidly as the number of variables is increased, and model flexibility is increased, but variance increases if we make our methods more flexible. The squared terms in MSE leads to a classic U-shape, and the optimum prediction will occur at some minimum
error point, that will be balanced between low bias and low variance contributions to the mean squared error (MSE).

The lasso method, introduced by Tibshirani [36], has become an increasingly popular method for feature selection. The lasso method belongs to a general class of penalized regression methods, also known as regularization or shrinkage methods. The idea of penalized regression was originally introduced in 1970 by Hoerl and Kennard [19], in order to allow estimation of ill-conditioned matrices in linear regression, by adding a penalty for the total sum of standardized coefficients to the sum of the squared squares metric. The penalty effectively budgets the total standardized coefficients: ridge regression can give improved predictions over least squares, because the decreased flexibility of the model decreases variance while adding bias. The lasso builds on this idea by introducing a different penalty, based on the sum of the absolute value of the coefficients.

All of these approaches can be generally stated using the following likelihood function $L$, which is a function of parameters $\lambda_1$, $\lambda_2$, and the vector $\beta$:

$$L(\lambda_1, \lambda_2, \beta) = |Y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|_1$$

where the ridge and lasso penalty terms are adjusted by $\lambda_1$ and $\lambda_2$, respectively:

$$|\beta|^2 = \sum_{j=1}^{p} \beta_j^2, \quad \text{and} \quad |\beta|_1 = \sum_{j=1}^{p} |\beta_j|$$

Letting $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$, the estimator for the coefficient vector $\beta$ can then be stated as the following optimization problem:

$$\min_{\beta} |Y - X\beta|^2, \quad \text{subject to} \ (1 - \alpha)|\beta|_1 + \alpha|\beta|^2 \leq t \text{ for some } t. \quad (6)$$

The constraint $(1 - \alpha)|\beta|_1 + \alpha|\beta|^2$ is called the elastic net penalty. If $\alpha = 0$, this becomes the lasso penalty; if $\alpha = 1$, this becomes the ridge penalty. Closed form solutions can be obtained for OLS and ridge regression. For the lasso and elastic net, estimation of the coefficient vector $\beta$ does not have a closed form solution, because of the nonlinearities introduced by the absolute value sign, and must be found by optimization. While ridge, lasso and the elastic net, do not dominate one another in all circumstances, it is easy to vary the parameters and check using cross-validation, in order to choose the best prediction method. For example, the elastic-net has been shown to be most effective in reducing the effects of multicollinearity.

The lasso is of particular interest for us because of its properties in variable selection. The
variable selection properties of the lasso method are well-documented, including its power in selecting variables among noisy datasets [extensive literature cited in 16, page 91]. The lasso method can also be extended to continuous and categorical variables using grouped variable approaches, any by imposing additional constraints on using lasso methods when seeking to identify possible interactions. Lim and Hastie [27] introduce hierarchical group-lasso regularization (HGLR), an approach to allow variables to be in more than one group. This allows for the inclusion of both continuous and categorical variables, as well as all of the pairwise interactions between them, to be included as overlapping group lasso problems, with an enforced hierarchy in the interactions only between the categorical levels.

For HGLR, the model for the vector of quantitative response \( Y \) is:

\[
Y = \mu + \sum_{i=1}^{p} X_i \theta_i + \sum_{i < j} X_{i:j} \theta_{i:j} + \epsilon
\]  

(7)

with intercept \( \mu \), predictor matrix \( X_i \) with \( p \) features as either categorical and continuous variables, and interactions between variables denoted by \( X_{i:j} \). The loss function for the squared error loss is given by:

\[
L(Y, X_{i;i \leq p}, X_{i:j}; \mu, \theta) = \frac{1}{2} \left\| Y - \mu \cdot 1 + \sum_{i=1}^{p} X_i \theta_i + \sum_{i < j} X_{i:j} \theta_{i:j} \right\|^2
\]  

(8)

The optimization problem for the group-lasso obtains estimates as the solution to:

\[
\min_{\mu, \beta} \frac{1}{2} \left\| Y - \mu \cdot 1 - \sum_{j=1}^{p} X^j \beta^j \right\|^2 + \lambda \sum_{j=1}^{p} \gamma_j |\beta_j|_2
\]  

(9)

where \( \lambda \) and the \( \gamma_j \) are used to penalize all of the coefficients and separate groups differently. Different groups of main effects and interactions can be added to the group-lasso problem to obtain the hierarchical interaction model. Proofs of how this condition enforces hierarchy and algorithmic details are presented in Lim and Hastie [27].

4.2.3. Cluster-wise regression

Most clustering problems are considered to be unsupervised problems, in that the true class of any particular observation may not be known. However, since clustering is frequently used to identify subgroups within the overall population, and then the clusters are then used to improve regression results, this becomes a combination of supervised and unsupervised problems, in which clustering is used to get a particular prediction for energy consumption; this may be referred to as ‘semi-supervised’ clustering, where another metric or dataset is
used to judge the effectiveness of clustering. In some sense, we often want both explanation and prediction: clustering as an explanatory method, in order to identify key subgroups and archetype buildings to better predict energy consumption in subsequent regression models.

Cluster-wise regression (CLR henceforth) is intended to carry out simultaneously clustering of a dataset, and the fitting of regression models to each cluster. Brusco et al. [2] trace the development of cluster-wise regression as a combination of cluster analysis and regression to the work of Spáth [31, 32, 33]. Adaptations includes multiple response variables [7] and multi criterion optimization [3].

This paper refers the reader to Leisch [25] for a more detailed formulation, but the key elements of CLR models are finite mixture models with $K$ components with the following form:

$$h(y|x, \psi) = \sum_{k=1}^{K} \pi_k f(y|x; \theta_k)$$  \hfill (10)

$$\pi_k \geq 0, \quad \sum_{k=1}^{K} \pi_k = 1$$  \hfill (11)

where $y$ is the dependent variable with conditional density $h$, $x$ is the vector of independent variables, $\pi_k$ is the prior probability of component $k$, $\theta_k$ is the component specific parameter vector for the density function $f$, and $\psi = (\pi_1, \ldots, \pi_k, \theta'_1, \ldots, \theta'_K)'$ is the vector of all parameters.

The posterior probability that the observation $(x, y)$ belongs to class $j$ is given by:

$$P(j|x, y, \psi) = \frac{\pi_j f(y|x, \theta_j)}{\sum_k \pi_k f(y|x, \theta_k)}$$  \hfill (12)

The posterior probabilities are used to segment the data by assigning each observation to the class with maximum posterior probability. The functions $f(y|x, \theta_j)$ are referred to as the mixture components, and can be implemented in the form of Gaussian response, i.e., the typical linear regression problem in equation 1.

We can then estimate a log-likelihood function for a sample of observations and maximize this subject to constraints on the mixing proportions. The optimization problem is frequently solved using the E-M algorithm [6] and multiple restarts to find avoid local maxima and to find the global maximum.

The analysis below will compare the typical practice of fitting a two-stage model, that is, clustering first and then performing linear regression second, with results from cluster-wise regression using density-based mixture models. One caution with CLR is that there is still
the potential for overfitting in cluster-wise regression. Since much of the improvement in global fit comes from the initial partitioning of the dataset into clusters, it is possible to obtain spurious correlations within clusters [2].

4.3. Integrated approach

Based on the theory presented in the previous section, this section will now propose an integrated approach that at least selects important global features in prediction, and then uses these to identify the key features for cluster-wise regression. Figure 1 illustrates the key steps.

Shrinkage step. First, shrinkage methods, also known as regularization or penalized regression methods as discussed in Section 4.2.2 above, are first applied in order to reduce the large number of features into parsimonious models of building energy use. The general lasso model highlights the most significant coefficients and their related features, and eliminates the less significant features that do not add any predictive power. A particular extension, hierarchical group-lasso regularization [27], is used to enforce a strong hierarchy among the continuous and categorical variables and their interactions. Cross validation is used to select the appropriate number of predictors based on the mean squared error, and to achieve the appropriate balance between bias and variance in prediction.

Cluster-wise regression. Using the features selected that give the best global fit, cluster-wise regression using mixture models is then implemented. The posterior probabilities for each observation are obtained. Simulating on these probabilities gives the most likely clusters in the data, and allows predictions for the estimated mixture components, and therefore the mean-squared error to be calculated.

5. Application

5.1. Data

The dataset for this paper has been previously used in two previous papers [21, 22], and so it will only briefly be described here. A description of the summary statistics and cleaning process for the dataset can be found in Hsu [21].

Datasets for large multifamily buildings in New York City were assembled, because large buildings represent 48% of all primary energy use in New York City (compared to 17% transportation and 35% from small buildings), and multifamily and office buildings represent 87% of all gross floor area of all large buildings [4]. The datasets were assembled from multiple data sources, including the U.S. Environmental Protection Agency’s Portfolio Manager, the City of New York’s Primary Land Use Tax Lot Output (PLUTO) database, real estate
and financial information from the CoStar Group, and census tract level information from the U.S. Census, including information from the 2010 decadal Census, the 2011 American Housing Survey, and the 2013 American Community Survey (ACS).

The outcome variable to be modeled for each building is the center-normalized total site (metered) energy. To explain this, data about energy performance and specific end-use characteristics was obtained through the City of New York’s annual benchmarking ordinance for commercial and multifamily buildings, Local Law 84. The City of New York collects whole-property energy benchmarking data for commercial and multifamily properties over 4,645 square meters (50,000 square feet) using the U.S. Environmental Protection Agency’s Portfolio Manager interface, including annualized energy use by fuel type, the distribution of space uses within the buildings, and selected measurements of end-uses for certain facility types. Energy information for parcels was then joined to key property and construction characteristics contained in City of New York’s PLUTO database using unique parcel identification numbers. Furthermore, each of these parcels can be located within a unique census tract using the City’s parcel shape files and the U.S. Census TIGER shape files. U.S. Census data was therefore downloaded for the 2010 Census, 2011 American Housing Survey, and the 2013 American Community Survey from the American FactFinder website. The census information was then spatially joined to each parcel using ArcGIS and CoStar financial information for the individual buildings. After joining the parcels, this resulted in 5,638 uniquely identified buildings, with extensive energy, tax, census, and financial information.

The dataset was then limited to multifamily housing buildings as classified in the Portfolio Manager database, resulting in 4,072 multifamily buildings. This classification is defined as when more than 50% of the gross area is used for that space use. This was further limited to buildings that constitute more than 75% of multifamily housing, since this eliminated only 4% of each category, and in order to get only buildings typical of each category. This resulted in 3,941 multifamily housing buildings. After cleaning, the datasets for the analysis have 3,902 multifamily housing buildings and more than 7,000 possible features and interactions.

5.2. Methods

Shrinkage methods were first applied to get the best coefficients for the entire dataset. Hierarchical group-lasso regularization (HGLR) was re-applied to this dataset as in Hsu [21], but with a different outcome variable, total building energy use. The R statistical language, glmnet, and glinternet packages were all used to implement the analysis [12, 26, 29]. As in the previous paper, HGLR outperformed the lasso, ridge, and elastic net approaches in terms of mean-squared error in a model with just six variables, as shown in Figure 2. The key variables to be used for the cluster-wise regression and their global coefficients are reported
in Table 1.

Then, CLR was carried out using finite mixture models. Estimation occurs via the E-M algorithm in the flexmix software package [25, 15]. The prediction error of the cluster-wise regression was evaluated using 10-fold cross validation.

Comparing the effectiveness of the integrated cluster-wise regression with separate clustering and regression methods was also evaluated using cross-validation. First, the data was split into ten folds for cross-validation, with 90% of the data used for training and 10% used for testing. On the training sets, since K-means clusters requires the numbers of clusters to be specified, K-means clustering was then run, choosing over the range of one to ten clusters. For each of the chosen number of clusters in the training set, linear regression was then run on each cluster to identify key predictors. In order to measure the prediction error in cross-validation, the K-means cluster identification was extrapolated from the training set to the test set using the clue package [20], and then the mean squared error between the test value and the predicted test value was calculated using the model for each cluster.

The separate clustering and regression methods used were K-means clustering and linear regression from the base stats package in the R language, and model-based clustering using the Mclust package [10].

5.3. Results & discussion

In terms of our original analysis goals presented in Section 2, we started with an approach that was designed to reduce high-dimensional data appropriately to a more parsimonious form. Shrinkage or regularization was used as a method to screen or select the most relevant features for a "global" fit, and then cluster-wise regression was used in order to identify key determinants within each subgroup. The cross-validated prediction results in Table 1 show that the cluster-wise regression consistently outperforms fitted models at the global level— including the lasso, ridge, elastic net, and HGLR—as well as outperforming regression fits to subgroups using the two-stage clustering and regression methods.

The cross-validated prediction errors from each method are shown in Tables 2, 4, and 6. The cluster-wise regression approach clearly outperforms the HGLR approach to the global data, as well as both of the two-stage clustering and regression methods (using density-based models and K-means, respectively).

The explanatory and interpretive power of the model can be judged in terms of the subgroups that are found in the dataset. A key consideration whether the clusters are useful in policymaking is whether the clusters are stable with respect to minor perturbations or changes in the data. Tables 2, 3, and 5 shows the number of clusters found in the overall dataset in cluster-wise regression, K-means regression, and model-based clustering.
Cross-validated results for cluster-wise regression shows that it consistently converges on four clusters in this dataset, but with different numbers of observations in each cluster. However, a comparison has not been carried out to determine the stability of the clusters with respect to the different folds. While it is possible to “validate” the clusters themselves using various approaches [18, 37], these approaches have not yet been tested.

The cluster assignments seem to show that there is consistently a large “typical” group, two smaller subgroups of highly similar buildings that can be predicted with high accuracy, and a group of outliers. However, because cluster assignments or labels can be permuted with respect to one another, it is not always clear if we can predict which cluster can be labelled as typical or as belonging to a subgroup of interest, other than looking at the total number of observations per cluster. The outlier groups, however, can be easily identified by their large variance ($\sigma$) compared to the other clusters.

The key determinants of energy consumption are reported in Table 7, and show clearly that different factors influence the behavior of each subgroup. This information could be used to target interventions or policies to each cluster. The significance patterns also vary by group, but unfortunately, the significance tests cannot strictly be relied upon because the particular reported model was selected from information from the initial model runs on the basis of log-likelihood calculations and the Bayesian Information Criterion [see 15, 25, for discussion].

5.4. Limitations

One of the main limitations on this research is the fact that despite using cluster-wise regression to integrate the clustering and prediction steps, the initial feature selection was still carried out using a pre-screening step. Ideally, feature selection would also be carried out in an integrated approach with the subsequent clustering and prediction within subgroups. This would require a global criterion that can balance many different concerns, including the best overall fit, the best fit within clusters, stability within clusters, and the best overall predictive fit. Development of such an ambitious approach is outside of the scope of this paper, unfortunately, but the next section will discuss promising future directions for research.

**Sparse cluster-wise regression.** Sparse cluster-wise regression does not yet exist. Ideally, we would be able to recover a sparse model that reflects the structure of the data, can be flexibly applied to out-of-sample data to get good predictions, and which yields clusters that are stable. Some work has been done in feature selection in clustering, particularly in high-dimensional genomic datasets [40] and variable selection or pre-screening approaches [34], as this paper has sought to apply. Unfortunately, this author has not found any work in cluster-wise regression that also performs feature selection.
Regression trees. One alternative approach that this paper did not try was classification and regression trees, or CART for short. As the name implies, CART approaches are often used to classify or provide prediction, but not necessarily to do both simultaneously. Given the amount of previous work put into improving CART approaches through bagging, boosting, and random forest approaches, this may be worth further investigation.

Software considerations. Finally, as demonstrated by the application example and methods discussion, the analysis and implementation of multiple different methods was only possible using many different software packages in the R statistical language. Any future approaches will only be useful if they are implemented in software for others to use and modify for further research.

6. Conclusions

This paper began by discussing how big data requires new statistical techniques, particularly for the tall ($n$ large) and wide ($p$ large) aspects, and for the high-dimensional aspects, where $p \gg n$. Competing goals of appropriateness for high-dimensions, predictive power, explanatory and interpretive power, and parsimony were all discussed. Related work in building energy consumption was then discussed, and how clustering and regression are often used together. The theory section examined two key issues: first, the importance of feature selection, and second, how to integrate explanatory and prediction approaches. This theory was then applied to finding subgroups and predicting the energy consumption of New York City multifamily buildings. The results were discussed in terms of their predictive accuracy and explanatory power. Finally, some directions for future research into sparse cluster-wise regression were suggested but have not yet been explored.
7. References


List of Figures

1 Integrated Approach Steps. .......................................... 23
2 Cross-validation prediction error for shrinkage methods. ................................... 24
Figure 1: Integrated Approach Steps.
Figure 2: Cross-validation prediction error for shrinkage methods.
# List of Tables

1. Variables and interactions selected by HGLR for clustering. ......................................................... 26
2. Mean-Squared Prediction Error for Cluster-wise Regression. Cluster-wise regression consistently converges on four main clusters. The total number of observations in the four clusters vary because of sampling for the 10-fold cross validation. The fourth cluster seems to consist mainly of outliers, due to poor predictions and high variance in the density mixture. ................................................................. 27
3. Number of Clusters Found in Density-Based Clustering. ............................................................ 28
4. CV Error Found in Density-Based Clustering with Subsequent Linear Regression. ................................................................. 29
5. Mean and Standard Deviation of Number of Observations Per Cluster for $K$-means. Number of clusters varies from 1 to 10 because $K$-means requires initial specification of total number of clusters. Means and standard deviations are averaged over 10-folds. ................................................................. 30
6. Mean and Standard Deviation of CV Error for $K$-means Clustering with Subsequent Linear Regression. Number of clusters varies from 1 to 10 because $K$-means requires initial specification of total number of clusters. Means and standard deviations are averaged over 10-folds. ............................................. 31
7. Mixture Components, i.e., Regression Coefficients for Each Cluster. ........................................ 32
<table>
<thead>
<tr>
<th>Variable 1</th>
<th>Variable 2</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>PctElectricityUse</td>
<td></td>
<td>-2.23</td>
</tr>
<tr>
<td>PctSteamUse</td>
<td></td>
<td>4.56</td>
</tr>
<tr>
<td>PctHHFElectric</td>
<td></td>
<td>0.04</td>
</tr>
<tr>
<td>PctElectricityUse</td>
<td>PctSteamUse</td>
<td>-11.94</td>
</tr>
<tr>
<td>PctElectricityUse</td>
<td>LogAssessedTot</td>
<td>0.04</td>
</tr>
<tr>
<td>PctDistrictSteamUse</td>
<td>PctYSB198089</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 1: Variables and interactions selected by HGLR for clustering.
<table>
<thead>
<tr>
<th>Fold</th>
<th>Obs. in Cluster</th>
<th>CV Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1  2  3  4</td>
<td>1  2  3  4</td>
</tr>
<tr>
<td>1</td>
<td>2239 742 364 155</td>
<td>0.07 0.15 0.04 3.03</td>
</tr>
<tr>
<td>2</td>
<td>1974 1095 306 147</td>
<td>0.06 0.15 0.06 1.73</td>
</tr>
<tr>
<td>3</td>
<td>2508 579 304 158</td>
<td>0.12 0.06 0.04 3.04</td>
</tr>
<tr>
<td>4</td>
<td>2301 820 255 146</td>
<td>0.07 0.17 0.05 3.93</td>
</tr>
<tr>
<td>5</td>
<td>2209 601 590 128</td>
<td>0.08 0.04 0.18 6.32</td>
</tr>
<tr>
<td>6</td>
<td>2005 831 585 132</td>
<td>0.08 0.05 0.17 2.69</td>
</tr>
<tr>
<td>7</td>
<td>2204 652 610 127</td>
<td>0.07 0.04 0.20 4.60</td>
</tr>
<tr>
<td>8</td>
<td>1856 917 648 129</td>
<td>0.09 0.06 0.13 1.55</td>
</tr>
<tr>
<td>9</td>
<td>2314 544 541 130</td>
<td>0.09 0.05 0.19 2.22</td>
</tr>
<tr>
<td>10</td>
<td>2005 698 683 129</td>
<td>0.08 0.18 0.05 6.44</td>
</tr>
<tr>
<td>Mean</td>
<td>2162 748 489 138</td>
<td>0.08 0.09 0.11 3.56</td>
</tr>
<tr>
<td>SD</td>
<td>187 163 154 11</td>
<td>0.02 0.05 0.07 1.66</td>
</tr>
</tbody>
</table>

Table 2: Mean-Squared Prediction Error for Cluster-wise Regression. Cluster-wise regression consistently converges on four main clusters. The total number of observations in the four clusters vary because of sampling for the 10-fold cross validation. The fourth cluster seems to consist mainly of outliers, due to poor predictions and high variance in the density mixture.
<table>
<thead>
<tr>
<th>Fold</th>
<th>Obs. in Cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1562</td>
</tr>
<tr>
<td>2</td>
<td>1956</td>
</tr>
<tr>
<td>3</td>
<td>2041</td>
</tr>
<tr>
<td>4</td>
<td>1988</td>
</tr>
<tr>
<td>5</td>
<td>1852</td>
</tr>
<tr>
<td>6</td>
<td>1093</td>
</tr>
<tr>
<td>7</td>
<td>1831</td>
</tr>
<tr>
<td>8</td>
<td>1504</td>
</tr>
<tr>
<td>9</td>
<td>2151</td>
</tr>
<tr>
<td>10</td>
<td>1811</td>
</tr>
<tr>
<td>Mean</td>
<td>1779</td>
</tr>
<tr>
<td>SD</td>
<td>297</td>
</tr>
</tbody>
</table>

Table 3: Number of Clusters Found in Density-Based Clustering.
<table>
<thead>
<tr>
<th>Fold</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.41</td>
<td>0.72</td>
<td>0.79</td>
<td>24.31</td>
<td>12.58</td>
<td>1.76</td>
<td>2.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>0.90</td>
<td>0.49</td>
<td>0.58</td>
<td>23.26</td>
<td>10.70</td>
<td>2.38</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0.98</td>
<td>0.47</td>
<td>23.06</td>
<td>1.42</td>
<td>14.88</td>
<td>0.27</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>0.99</td>
<td>0.38</td>
<td>23.61</td>
<td>0.94</td>
<td>14.66</td>
<td>1.45</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>1.73</td>
<td>0.54</td>
<td>21.02</td>
<td>3.24</td>
<td>0.52</td>
<td>0.48</td>
<td>81.03</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>1.79</td>
<td>0.38</td>
<td>1.04</td>
<td>26.46</td>
<td>1.05</td>
<td>5.97</td>
<td>2.26</td>
<td>NA</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>0.81</td>
<td>0.42</td>
<td>1.04</td>
<td>24.33</td>
<td>13.82</td>
<td>9.98</td>
<td>2.84</td>
<td>0.94</td>
<td>0.03</td>
</tr>
<tr>
<td>8</td>
<td>1.82</td>
<td>0.44</td>
<td>24.09</td>
<td>1.83</td>
<td>0.55</td>
<td>0.61</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>9</td>
<td>0.64</td>
<td>3.07</td>
<td>1.91</td>
<td>25.78</td>
<td>31.63</td>
<td>21.02</td>
<td>0.19</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>10</td>
<td>1.49</td>
<td>0.48</td>
<td>23.73</td>
<td>1.28</td>
<td>0.35</td>
<td>4.59</td>
<td>0.09</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.26</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>12.09</td>
<td>13.28</td>
</tr>
<tr>
<td></td>
<td>10.07</td>
<td>4.85</td>
</tr>
<tr>
<td></td>
<td>8.84</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 4: CV Error Found in Density-Based Clustering with Subsequent Linear Regression.
<table>
<thead>
<tr>
<th>K</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3929.0</td>
<td>(0.0)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2920.4</td>
<td>1008.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(445.1)</td>
<td>(445.1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2614.2</td>
<td>992.7</td>
<td>322.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(136.8)</td>
<td>(147.4)</td>
<td>(114.4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2323.4</td>
<td>1085.5</td>
<td>344.3</td>
<td>175.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(81.4)</td>
<td>(20.2)</td>
<td>(25.2)</td>
<td>(37.3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2133.2</td>
<td>1134.0</td>
<td>330.7</td>
<td>187.7</td>
<td>143.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(59.8)</td>
<td>(6.5)</td>
<td>(2.4)</td>
<td>(24.0)</td>
<td>(47.5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1882.8</td>
<td>986.5</td>
<td>509.2</td>
<td>330.9</td>
<td>164.6</td>
<td>55.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(41.3)</td>
<td>(40.5)</td>
<td>(42.4)</td>
<td>(2.7)</td>
<td>(1.9)</td>
<td>(38.6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1773.1</td>
<td>886.0</td>
<td>511.3</td>
<td>328.2</td>
<td>258.0</td>
<td>137.7</td>
<td>34.7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(176.7)</td>
<td>(96.4)</td>
<td>(110.7)</td>
<td>(40.4)</td>
<td>(78.6)</td>
<td>(52.4)</td>
<td>(8.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1506.8</td>
<td>880.2</td>
<td>555.6</td>
<td>337.8</td>
<td>284.3</td>
<td>197.4</td>
<td>136.9</td>
<td>30.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(241.4)</td>
<td>(58.8)</td>
<td>(89.6)</td>
<td>(52.0)</td>
<td>(53.0)</td>
<td>(43.8)</td>
<td>(50.5)</td>
<td>(10.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1316.0</td>
<td>890.1</td>
<td>614.2</td>
<td>305.4</td>
<td>256.4</td>
<td>206.4</td>
<td>165.3</td>
<td>145.3</td>
<td>29.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(113.7)</td>
<td>(42.0)</td>
<td>(27.0)</td>
<td>(28.7)</td>
<td>(23.9)</td>
<td>(27.7)</td>
<td>(7.4)</td>
<td>(14.5)</td>
<td>(6.5)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1157.3</td>
<td>906.1</td>
<td>582.3</td>
<td>341.1</td>
<td>260.7</td>
<td>222.1</td>
<td>169.0</td>
<td>149.0</td>
<td>114.9</td>
<td>26.6</td>
</tr>
<tr>
<td></td>
<td>(121.5)</td>
<td>(37.7)</td>
<td>(61.1)</td>
<td>(77.2)</td>
<td>(35.9)</td>
<td>(29.8)</td>
<td>(17.1)</td>
<td>(10.0)</td>
<td>(32.0)</td>
<td>(9.9)</td>
</tr>
</tbody>
</table>

Table 5: Mean and Standard Deviation of Number of Observations Per Cluster for $K$-means. Number of clusters varies from 1 to 10 because $K$-means requires initial specification of total number of clusters. Means and standard deviations are averaged over 10-folds.
<table>
<thead>
<tr>
<th>K</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.32)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.23</td>
<td>8.99</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.72)</td>
<td>(4.83)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.90</td>
<td>3.71</td>
<td>15.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.22)</td>
<td>(2.16)</td>
<td>(7.10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.94</td>
<td>0.96</td>
<td>20.10</td>
<td>11.98</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.22)</td>
<td>(0.62)</td>
<td>(2.23)</td>
<td>(6.76)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.94</td>
<td>0.61</td>
<td>20.62</td>
<td>4.40</td>
<td>9.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.23)</td>
<td>(0.22)</td>
<td>(1.44)</td>
<td>(6.64)</td>
<td>(7.15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.98</td>
<td>0.61</td>
<td>168.94</td>
<td>20.63</td>
<td>11.84</td>
<td>3.09</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.25)</td>
<td>(0.21)</td>
<td>(531.81)</td>
<td>(1.40)</td>
<td>(6.95)</td>
<td>(3.87)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.99</td>
<td>0.61</td>
<td>0.79</td>
<td>10.89</td>
<td>12.48</td>
<td>11.16</td>
<td>21.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.25)</td>
<td>(0.22)</td>
<td>(0.30)</td>
<td>(10.47)</td>
<td>(10.50)</td>
<td>(7.41)</td>
<td>(41.49)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.00</td>
<td>0.69</td>
<td>0.71</td>
<td>7.52</td>
<td>16.77</td>
<td>5.36</td>
<td>9.18</td>
<td>NA</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.29)</td>
<td>(0.18)</td>
<td>(0.56)</td>
<td>(10.73)</td>
<td>(8.70)</td>
<td>(5.65)</td>
<td>(8.14)</td>
<td>(NA)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.07</td>
<td>0.63</td>
<td>0.44</td>
<td>3.23</td>
<td>7.27</td>
<td>15.61</td>
<td>9.54</td>
<td>9.89</td>
<td>6.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.32)</td>
<td>(0.16)</td>
<td>(0.13)</td>
<td>(6.84)</td>
<td>(9.92)</td>
<td>(11.22)</td>
<td>(7.98)</td>
<td>(6.51)</td>
<td>(16.28)</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1.09</td>
<td>0.73</td>
<td>0.44</td>
<td>0.75</td>
<td>7.36</td>
<td>14.18</td>
<td>13.00</td>
<td>9.68</td>
<td>5.05</td>
<td>NA</td>
</tr>
<tr>
<td></td>
<td>(0.34)</td>
<td>(0.17)</td>
<td>(0.16)</td>
<td>(0.38)</td>
<td>(10.84)</td>
<td>(10.54)</td>
<td>(8.47)</td>
<td>(7.06)</td>
<td>(5.05)</td>
<td>(NA)</td>
</tr>
</tbody>
</table>

Table 6: Mean and Standard Deviation of CV Error for $K$-means Clustering with Subsequent Linear Regression. Number of clusters varies from 1 to 10 because $K$-means requires initial specification of total number of clusters. Means and standard deviations are averaged over 10-folds.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Comp. 1</th>
<th>Comp. 2</th>
<th>Comp. 3</th>
<th>Comp. 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.033</td>
<td>-0.419</td>
<td>0.104</td>
<td>-0.044</td>
</tr>
<tr>
<td>PctElectricityUse</td>
<td>-3.245</td>
<td>-4.001</td>
<td>0.043</td>
<td>-1.617</td>
</tr>
<tr>
<td>PctDistrictSteamUse</td>
<td>0.291</td>
<td>0.843</td>
<td>1.121</td>
<td>-0.182</td>
</tr>
<tr>
<td>LogAssessTot</td>
<td>-0.364</td>
<td>-0.397</td>
<td>0.146</td>
<td>-0.058</td>
</tr>
<tr>
<td>PctYSB198089</td>
<td>0.015</td>
<td>0.338</td>
<td>0.055</td>
<td>-0.016</td>
</tr>
<tr>
<td>PctHHFElectric</td>
<td>-0.005</td>
<td>0.152</td>
<td>0.015</td>
<td>0.079</td>
</tr>
<tr>
<td>PctElectricityUseXPctDistrictSteamUse</td>
<td>-0.253</td>
<td>-0.662</td>
<td>-1.061</td>
<td>0.200</td>
</tr>
<tr>
<td>PctElectricityUseXLogAssessTot</td>
<td>3.312</td>
<td>3.329</td>
<td>-0.372</td>
<td>1.027</td>
</tr>
<tr>
<td>PctDistrictSteamUseXPctYSB198089</td>
<td>-0.004</td>
<td>0.121</td>
<td>-0.017</td>
<td>-0.001</td>
</tr>
</tbody>
</table>

Table 7: Mixture Components, i.e., Regression Coefficients for Each Cluster.
Planning in the Big Data Era: A Theme Study of Residential Energy Policy

Hossein Estiri, PhD
Institute of Translational Health Sciences
University of Washington
hestiri@uw.edu
1. Abstract

This paper re-conceptualizes the planning process in the big data era based on the improvements that non-linear modeling approaches provide over the mainstream linear approaches. First, it demonstrates challenges of conventional linear methodologies in modeling complexities of residential energy use, addressing the “variety” from the three Vs of big data. Suggesting a non-linear modeling schema to analyze household energy use, the paper develops its discussion around the repercussions of the use of non-linear modeling in energy policy and planning. Planners / policy-makers are not often equipped with the tools needed to translate complex scientific outcomes into policies. To fill this gap, this work proposes modifications in the traditional planning process in order to be able to benefit from the abundance of data and the advances in analytical methodologies. The conclusion section introduces three short-term repercussions of this work for energy policy (and planning, in general) in the big data era: tool development, data infrastructures, and planning education.

Keywords: energy policy; residential buildings; non-linear modeling; big data; planning process.

2. Introduction

According to the International Energy Outlook 2013 (IEO2013), by 2040, world energy consumption will be 56% higher than its 2010 level, most of which is due to socioeconomic transformations in developing countries (U.S. Energy Information Administration, 2013b). This increase is expected to occur despite the existence of several global agreements within the past few decades on significantly reducing greenhouse gases (GHGs) and energy consumption (e.g. the Kyoto Protocol, adopted in December 1997 and entered into force in February 2005).

Globally, buildings (residential and commercial) consume between 20% and 40% of total energy (Norman, MacLean, & Kennedy, 2006; Roaf, Crichton, & Nicol, 2004; Swan & Ugursal, 2009). About 20% to 30% of the total energy demand is for residential use. In 2013, 22% of the energy consumed and 21% of the CO₂ emissions produced in the U.S. came from the residential sector
(Figure 1 and Figure 2), both of which are expected to slightly diminish in their share to 20% and 19%, respectively, due to faster increases in industrial and commercial energy consumption (U.S. Energy Information Administration, 2013a). Technological improvements are expected to diminish growth rates in residential and transportation energy use. Since developed countries have greater access to up-to-date technologies, energy consumed in the residential buildings is likely to increase at a slower pace in developed counties, with an average of 14% in developed and 109% in developing countries (Figure 3) (U.S. Energy Information Administration, 2013b).

Figure 1. U.S. Energy Consumption by Sector, 2013, 2020, 2030, and 2040. Data source: (U.S. Energy Information Administration, 2013a)
Nevertheless, to many consumers (households), researchers, and policymakers, the energy consumed at homes has become an invisible resource (Brandon & Lewis, 1999). A clear understanding of residential energy consumption is the key constituent of effective energy policy and planning (Brounen, Kok, & Quigley, 2012; Hirst, 1980). Two main reasons explain the uncertainties in household energy consumption research and theory, obstructing the clear understanding needed for effective energy policy. First, conventional research has commonly used linear methodologies to analyze energy use in the residential sector, failing to account for its complexities. Second, there is a lack of publicly available energy use data, which has intensified the methodological issues in studying residential energy consumption.

3. Problem: Prior research underestimated the human role

Due to its complexities, investigating the policy implications of behavioral determinants of residential energy consumption has received little attention in prior research (Brounen et al., 2012). Traditionally, the debate on residential energy conservation has neglected the role of occupants’ behaviors by excessively focusing on technical and physical attributes of the housing unit (Brounen et al., 2012; Kavgic et al., 2010; Kriström, 2006; Lutzenhisler, 1993). Since the early 1990s, energy research and policy have primarily concentrated either on the supply of energy or the efficiency of buildings, neglecting social and behavioral implications of energy demand (Aune, 2007; Brounen et al., 2012; Lutzenhisler, 1992, 1994; Pérez-Lombard, Ortiz, & Pout, 2008).
Engineering and economic approaches underestimate the significance of occupant lifestyles and behaviors (Lutzenhiser, 1992).

“Engineers and other natural scientists continue to usefully develop innovative solutions to the question of “how we can be more efficient?” However their work does not answer the question “why are we not more energy-efficient, when clearly it is technically possible for us to be so?” (Crosbie, 2006)

In most energy demand studies, only a limited set of socio-demographic attributes are involved (O’Neill & Chen, 2002), due to methodological or data deficiencies. Moreover, the complexity of the human role in the energy consumption process makes meaningful interpretation of modeling results rather difficult, which in turn leads to ambiguities and a limited understanding of the role of socioeconomic and behavioral determinants of residential energy use. For example, Yu et al. (2011) suggest that because the influence of socioeconomic factors on energy consumption are reflected in the effect of occupant behaviors, “there is no need to take them into consideration when identifying the effects of influencing factors” (Yu, Fung, Haghighat, Yoshino, & Morofsky, 2011, p. 1409). Whereas, buildings do not consume energy, per se, and residential energy demand is driven by human activity.

4. Why has the role of human been underestimated?

4.1. Linearity vs. Non-linearity

Understanding and theorizing household energy use processes and repercussions are “a far from straightforward matter” (Lutzenhiser, 1997, p. 77).

“Household energy consumption is not a physics problem, e.g., with stable principles across time and place, conditions that can be clearly articulated, and laboratory experiments that readily apply to real world.” (Moezzi & Lutzenhiser, 2010, p. 209)

Linear analytical methodologies have been a research standard in understanding domestic energy consumption. The assumption of linearity (where the dependent variable is a linear function of independent variables) and the difficulty to ascertain any causal interpretations (i.e. the correlation vs. causation dilemma) are major downsides of traditional methodologies, such as ordinary multivariate regression models (Kelly, 2011). As a consequence of the predominant assumption of linearity in energy consumption research, “the present [conventional] energy policy still conveys
a ‘linear’ understanding of the implementation of technology” (Aune, 2007, p. 5463), while linear models cannot explain the complexities of household-level energy consumption (Kelly, 2011). For better energy policies, a better understanding of the complexities of its use is needed (Aune, 2007; Hirst, 1980; Swan & Ugursal, 2009).

4.2. Lack of publicly available data

A major problem in residential energy consumption research is that “the data do not stand up to close scrutiny” (Kriström, 2006, p. 96). Methodological approaches lag behind theoretical advances, partly because data used for quantitative analysis often do not include the necessary socio-demographic, cultural, and economic information (Crosbie, 2006). In addition, the absence of publicly available high-resolution energy consumption data has hindered development of effective energy research and policy (Hirst, 1980; Kavgic et al., 2010; Lutzenhiser, Moezzi, Hungerford, & Friedmann, 2010; Min, Hausfather, & Lin, 2010; Pérez-Lombard et al., 2008).

Even though relevant data are being regularly collected by different organizations, such data sources do not often become publicly known (Hirst, 1980). Conventional wisdom and modeling practices of energy consumption are often based on “averages” derived from aggregated data (e.g. average energy consumption of an appliance, a housing type, a car, etc.), which do not explicitly reflect human choice of housing and other energy consumptive goods (Lutzenhiser & Lutzenhiser, 2006).

5. Non-linear Modeling

Like most urban phenomena, residential energy use is an “outcome” of a set of complex interactions between multiple physical and behavioral factors. Figure 4 illustrates one dimension of the difference between linear and non-linear approaches. A linear approach often considers the outcome as a “dependent” variable that correlates with a set of “independent” variables, which in turn, may correlate with each other, as well. Clear examples of linear models are various type of multivariate regression models. In a non-linear approach, however, the outcome is the result of a set of cause-and-effect interactions between the predictor variables. This means that if one of the predictor variables changes, it will be unrealistic to assume that other variables would hold constant (a “gold standard” in reporting regression results) – with the exception of totally exogenous variables.
Figure 4. Comparing linear and non-linear modeling approaches

This difference in the two approaches can be game changing, as the non-linear approach can reveal an often hidden facet of effects on the outcome, the “indirect” effects. Research has shown that, for example, linear approaches significantly underestimate the role of household characteristics on energy use in residential buildings, as compared with the role of housing characteristics (Estiri, 2014a, 2014b). This underestimation has formed the conventional understanding on residential energy and guided current policies that are “too” focused on improving buildings’ energy efficiency.

Figure 5 illustrates a non-linear conceptualization of the energy consumption at the residential sector. According to the figure, households have a direct effect on energy use through their appliance use behaviors. Housing characteristics, such as size, quality, and density also influence energy use directly. Household characteristics, however, influence the characteristics of the housing unit significantly – which is labeled as housing choice. In addition to their direct effect, through the housing choice, households have an indirect effect on energy consumption, which has been dismissed with the use of linear methodologies, and so, overlooked in conventional thinking and current policies.
Figure 5. A non-linear conceptual model of the impact of the household and the housing unit on energy consumption. Source: (Estiri, 2014a).

6. A proposed non-linear modeling schema

Energy use in the residential sector is a function of local climate, the housing unit, energy markets, and household characteristics and behaviors. A conventional linear approach to household energy use correlates all of the predictors to the dependent variable (Figure 6). Figure 7, instead, illustrates a non-linear model that incorporates multiple interactions between individual determinants of energy consumption at the residential sector. Results of the non-linear model will be of more use for energy policy.

Figure 6. Graphical model based on the linear approach. All predictors correlate with the dependent variable, while mediations and interactions among variables are neglected.
The recommended graphical model (Figure 7) can be operationalized in form of 10 simultaneous equations – with 69 parameters to be estimated:

### Energy Use
\[
E_{\text{Energy Use}} = \beta_{14} \text{Age} + \beta_{15} \text{Gender} + \beta_{17} \text{Race/Ethnicity} + \beta_{18} \text{Income} + \beta_{19} \text{Education} + \beta_{20} \text{Household Size} + \beta_{21} \text{Marital Status} + \beta_{22} \text{Housing Type} + \beta_{23} \text{Tenure Type} + \beta_{24} \text{Energy Price} + \beta_{25} \text{Local Climate} + \epsilon_1
\]

### Income
\[
\text{Income} = \beta_{31} \text{Age} + \beta_{32} \text{Gender} + \beta_{33} \text{Race/Ethnicity} + \beta_{34} \text{Education} + \beta_{35} \text{Marital Status} + \epsilon_2
\]

### Education
\[
\text{Education} = \beta_{36} \text{Age} + \beta_{37} \text{Gender} + \beta_{38} \text{Race/Ethnicity} + \epsilon_3
\]

### Household Size
\[
\text{Household Size} = \beta_{39} \text{Age} + \beta_{40} \text{Gender} + \beta_{41} \text{Race/Ethnicity} + \beta_{42} \text{Income} + \beta_{43} \text{Education} + \beta_{44} \text{Marital Status} + \epsilon_4
\]

### Marital Status
\[
\text{Marital Status} = \beta_{45} \text{Age} + \beta_{46} \text{Gender} + \beta_{47} \text{Education} + \epsilon_5
\]

### Housing Age
\[
\text{Housing Age} = \beta_{48} \text{Age} + \beta_{49} \text{Race/Ethnicity} + \beta_{50} \text{Income} + \beta_{51} \text{Education} + \epsilon_6
\]

### Housing Size
\[
\text{Housing Size} = \beta_{52} \text{Age} + \beta_{53} \text{Race/Ethnicity} + \beta_{54} \text{Income} + \beta_{55} \text{Education} + \beta_{56} \text{Household Size} + \beta_{57} \text{Marital Status} + \beta_{58} \text{Housing Type} + \epsilon_7
\]

### Housing Type
\[
\text{Housing Type} = \beta_{59} \text{Age} + \beta_{60} \text{Race/Ethnicity} + \beta_{61} \text{Income} + \beta_{62} \text{Education} + \beta_{63} \text{Marital Status} + \beta_{64} \text{Housing Size} + \beta_{65} \text{Tenure Type} + \epsilon_8
\]

### Tenure Type
\[
\text{Tenure Type} = \beta_{66} \text{Age} + \beta_{67} \text{Race/Ethnicity} + \beta_{68} \text{Income} + \beta_{69} \text{Marital Status} + \beta_{70} \text{Housing Size} + \epsilon_9
\]

### Housing Quality
\[
\text{Housing Quality} = \beta_{71} \text{Age} + \beta_{72} \text{Race/Ethnicity} + \beta_{73} \text{Income} + \beta_{74} \text{Education} + \beta_{75} \text{Energy Price} + \epsilon_{10}
\]

There are five exogenous variables in this model: age, gender, race/ethnicity, local climate, and energy price. All housing-related characteristics can be predicted with household characteristics (which can be improved by adding other influential variables). The parameters in these simultaneous equations can be estimated using a variety of software packages. How the estimated parameters can be used in planning and policy is yet another challenge.
7. Scientists, planners, and complex modeling outcomes

“For the theory-practice iteration to work, the scientist must be, as it were, mentally ambidextrous; fascinated equally on the one hand by possible meanings, theories, and tentative models to be induced from data and the practical reality of the real world, and on the other with the factual implications deducible from tentative theories, models and hypotheses.” (Box, 1976, p. 792)

The better we – as individuals, planners, policy-makers – process complexities, the better decisions we’ll make. Future policies need to be smarter by taking more complexities into account. With the current growing computational capacities, it is quite feasible to estimate such complex models – models can be connected and estimated using live data, as well. Further, modern analytical algorithms can easily handle more complex models (models with increasing number or parameters). Clearly, we won’t be short of tools and technologies to model more and more complexities.

However, as the models get more complicated – and ideally produce more realistic explanations for energy consumption – translation of their results for policy and planning will become harder. Planners and policy-makers are not equipped with the required skillset to understand and interpret sophisticated modeling outcomes. Their strengths are, in turn, in developing policies and plans that operationalize community goals. I suggest, in the big data era, planning can benefit from the abundance of data – of varying types – and the advances in computational and analytical techniques through a planning process that is accordingly modified.

8. A modified planning process

The traditional planning process is not capable of directly incorporating complex scientific outcomes into policy development. The three primary steps in traditional planning process are: (1) gathering data; (2) transforming data into information; and (3) setting goals and objectives. Policies often follow explicit goals arrived at as the fourth step in the traditional planning process. There seems to be a missing link to connect complex modeling outcomes with the production of policy; perhaps an interface that can help planners and policy-makers set explicit goals for their respective communities.
The planning process needs modification to adapt to and benefit from this new Big Data era, with the abundance of data and growing advances in computer analytics. What is required for the outcomes of advanced complex modeling to be used in planning and policy is a paradigm shift in planning practice: a modified planning process (Figure 8).

As I mentioned earlier, the traditional planning process often begins with data gathering. I also discussed that data unavailability is an important issue that has hindered the advancement of residential energy consumption research and policy. Local utility companies are concerned about privacy issues. In addition, energy data needs to be connected to population, market, and climate data in a standardized way, to become useful for research and policy purposes.

The first step in this proposed planning process is a data collection and integration infrastructure comprised of energy, population, market, regulations, and climate data. There are various examples of federated data sharing infrastructures in health sciences that were developed using appropriate data governance and information architecture. Given that the bars for privacy are often set very high for health data, it should be feasible to develop similar data infrastructures for energy policy and research. Establishment of such integrated data infrastructures will require both technical and human components. Clearly, we will be needing data centers that can host the data, as well as cloud-based data sharing and querying technologies. But, technologies are only useful...
once the data is available – the foundation for data collection and integration are built. Here is where human role becomes important. To build a consensus among the data owners (utility companies, households, government or local agencies) multiple rounds of negotiations are conceivable. There also needs to be proper data governance in place before data can be collected, integrated, and shared with policy researchers.

New technologies (e.g., cloud computing, etc.) have made it easier to share and store data. Computer processing and analytics are also advancing rapidly, making it possible to process more data and complexities, faster and more efficiently (in its statistical denotation). There are several modern analytical approaches that can analyze more complexities, and can provide simulations. I suggest that the traditional analysis in planning process (step 2) should be enhanced/replaced by incorporating advanced modeling algorithms that are trainable and connected to live data. This process involves scientific discoveries.

Yet, planners and policy-makers should not be expected to be able to utilize complex modeling results directly into planning and policy-making. The findings of such analyses and simulations need to be made explicit via a policy interface. Using the policy interface, planners and policy-makers would be able: (1) to explicitly monitor the effects of various variables on energy consumption and results of a simulated intervention, and (2) to modify the analytical algorithms, if needed, to improve the outcomes. The interface should provide explicit goals for planners and policy-makers, making it easier to reach conclusions and assumptions.

From the explicit goals, designing smart policies is only a function of the planners’ / policy-makers’ innovativeness in finding the best ways (i.e., smartest policies) for their respective localities to achieve their goals. Smart policies are context-dependent and need to be designed in close cooperation with local stakeholders, as all “good” policies are supposed to. For example, if reducing the impact of income on housing size by X% was the goal, then changes in property taxes might be the best option in one region, while in another region changes in design codes could be the solution. Once smart policies are implemented, the results will be captured in the data infrastructure and used for further re-iterations of the planning process.

Conclusion

This study built upon a new approach to energy policy research: accounting for more complexities of the energy consumption process can improve conventional understanding and
produce results that are useful for policy. I suggested that in order for planner and policy makers to benefit from incorporation of complex modeling practices and the abundance of data, modifications are essential in the traditional planning process. More elaborations around the proposed modified planning process will require further work and collaborations within the urban planning-big data community. Regarding the modified planning process, in the short-run, three areas of further research can be highlighted.

First is developing prototype policy interfaces. The non-linear modeling that I proposed in this work can be operationalized and estimated using a variety of software packages. More important, however, is the integration of the proposed non-linear model into the corresponding policy interface. Energy Policy Analytics Dashboard (E-PAD) is a work-in-progress of the author towards this goal. More work needs to be done in this area using different methodologies, as well as developing more complex algorithms to understand more of the complexities in energy use in the residential sector – and perhaps, in other sectors.

Without integrated data it will be impossible to understand the complexities of energy use patterns – or any other urban phenomenon. Therefore, it is important to invest on city- and/or region-wide initiatives to securely collect and integrate data from different organizations. As the second area of future work, although establishing such initiatives and preparing the required socio-technical data infrastructure may not be a direct task for planners (for the time being), it certainly will be within the scope of work for local governments and planning / urban studies scholars.

Finally, the proposed modifications to planning process has important implications for planning education. It will be crucial for planning practitioners or scholars in Big Data era to be able to effectively play role across one or more steps of the proposed planning process. When there is abundance of data, planning education needs to incorporate more hands on methodological training for planners in order to familiarize them with [at least] basic concepts of using data and data interfaces smartly. There also needs to be training around developing data architectures and infrastructures, especially for planning scholars to integrate urban data. Training options will also be helpful for planners to understand the required governance and negotiations related obtaining and maintenance of data.
References


Appendix A

Workshop Program Committee

Conference Organizers

Piyushimita (Vonu) Thakuriah, University of Glasgow, Conference Chair
Nebiyou Tilahun, University of Illinois at Chicago, Conference Co-Chair
Moira Zellner, University of Illinois at Chicago, Conference Co-Chair

Program Committee Members

Jochen Albrecht, Hunter College, New York, USA
Budhudendra Bhaduri, Oak Ridge National Laboratory, USA
Max Dieber, University of Illinois at Chicago, USA
Lise Dirks, University of Illinois at Chicago, USA
Yingling Fan, University of Minnesota, USA
Joseph Ferriera, MIT, USA
Bernard Fingleton, University of Cambridge, UK
Steven French, Georgia Tech, USA
Glenn Geers, NICTA, Australia
Joe Grengs, University of Michigan, USA
Richard Harris, University of Bristol, UK
Andy Hudson-Smith, University College London, USA
Bin Jiang, University of Gavle, Sweden
Timothy Johnson, University of Illinois at Chicago, USA
David Levinson, University of Minnesota, USA
Jane Lin, University of Illinois at Chicago, USA
Geoff Meen, University of Reading, UK
Harvey Miller, Ohio State University, USA
Scott Page, University of Michigan, USA
Alenka Poplin, HefenCity Universitat Hamburg, Germany
Gwilym Pryce, University of Glasgow, UK
Laxmi Ramasubramanium, Hunter College, New York, USA
Piyushimita(Vonu) Thakuriah, University of Glasgow, UK
Nebiyou Tilahun, University of Illinois at Chicago, USA
Shashi Shekhar, University of Minnesota, USA
Paul Waddell, University of California, Berkeley, USA
Shaowen Wang, University of Illinois Urbana-Champaign, USA
Ouri Wolfson, University of Illinois at Chicago, USA
Jeremy Wu, George Washington University, USA
Sue Zielinski, University of Michigan, USA
Moira Zellner, University of Illinois at Chicago, USA
Appendix B

Participants of Workshop 1

University of Glasgow, UK
Piyushimita (Vonu) Thakuriah, Workshop Chair, Department of Urban Studies and School of Engineering

University of Illinois at Chicago
Nebiyou Tilahun, Co-Chair, Department of Urban Planning and Policy
Moira Zellner, Co-Chair, Department of Urban Planning and Policy
William “Max” Dieber, Urban Data Visualization Lab
Lisa Dirks, Urban Transportation Center
Timothy Johnson, Survey Research Lab
Jane Lin, Department of Civil and Materials Engineering

University of Utah
Harvey Miller, Department of Geography

National ICT of Australia, Sydney, Australia
Glenn Geers, Infrastructure Transport & Logistics

University of Minnesota
Yingling Fan, Humphrey Institute of Public Policy
David Levinson, Department of Civil Engineering
Shashi Shekhar, Department of Computer Science

Oak Ridge National Labs
Budhundendra Bhaduri, Geographic Information Science and Technology Group

Hunter College
Jochen Albrecht, Department of Geography
Laxmi Ramasubramanian, Urban Affairs and Planning

University of Michigan
Joe Grengs, Urban and Regional Planning, University of Michigan
Scott Page, Departments of Economics and Political Science
Sue Zielinski, Project SMART, University of Michigan

Georgia Institute of Technology
Steven French, School of City and Regional Planning

Private consultants
Jeremy Wu, (formerly at the U.S. Census Bureau), Washington, D.C.
Zipei Tu, Director of Asia Pacific Area, KIT Solutions, LLC
### Author Index

- Adlakha, Deepti 638
- Afzalan, Nader 294
- Aifandopoulou, Georgia 514
- Alruheil, Anna 313
- Anderson, Roger 485
- Antunes, Francisco 276
- Arcaute, Elsa 780
- Arribas-Bel, Daniel 1275
- Bakens, Jessie 1275
- Barchers, Camille 345
- Batty, Michael 712, 780
- Ben-Elia, Eran 1188
- Benenson, Itzhak 1188
- Bhaduri, Budhendra 1027
- Bigazzi, Alexander 553
- Biging, Gregory 313
- Boundy, Emma 896
- Cabanas, Orlando 734
- Carr, Sara 677
- Castro, Miguel 608
- Cervone, Guido 1013
- Cheng, Qimin 699
- Cheshire, James 712
- Chiu, Yi-Chang 528
- Chow, Joseph 463
- Cole, Sam 1258
- Cottrill, Caitlin 226
- Davidson, Adam 941
- Delaney, Phil 756
- Desouza, Kevin 379
- Dirks, Lise 398
- Donegan, Mary 575
- Dong, Jing 332
- Dunks, Richard 608
- Dutta, Promiti 485
- Elazzabi, Basem 808
- Erhardt, Gregory 780
- Estiri, Hossein 1327
- Evans, Christian 665
<table>
<thead>
<tr>
<th>Name</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kotz, Andrew</td>
<td>360</td>
</tr>
<tr>
<td>Kvan, Tom</td>
<td>756</td>
</tr>
<tr>
<td>Lang, Wei</td>
<td>313</td>
</tr>
<tr>
<td>Lassiter, Allison</td>
<td>677</td>
</tr>
<tr>
<td>Levinson, David</td>
<td>1209</td>
</tr>
<tr>
<td>Li, Pengfei</td>
<td>332</td>
</tr>
<tr>
<td>Lichter, Michal</td>
<td>979</td>
</tr>
<tr>
<td>Lim, Jeongki</td>
<td>608</td>
</tr>
<tr>
<td>Liu, Cheng</td>
<td>1027</td>
</tr>
<tr>
<td>Liu, Tianlan</td>
<td>33</td>
</tr>
<tr>
<td>Liu, Xintao</td>
<td>463</td>
</tr>
<tr>
<td>Lock, Oliver</td>
<td>780</td>
</tr>
<tr>
<td>Long, Ying</td>
<td>1173</td>
</tr>
<tr>
<td>Lu, Wei</td>
<td>1027, 1050</td>
</tr>
<tr>
<td>Ma, Yu-Luen</td>
<td>528</td>
</tr>
<tr>
<td>Machemehl, Randy</td>
<td>498</td>
</tr>
<tr>
<td>Maier, David</td>
<td>808</td>
</tr>
<tr>
<td>Maier, William F.</td>
<td>313</td>
</tr>
<tr>
<td>Manduca, Robert</td>
<td>49</td>
</tr>
<tr>
<td>Megler, Veronika</td>
<td>808</td>
</tr>
<tr>
<td>Minsker, Barbara</td>
<td>846</td>
</tr>
<tr>
<td>Miranda-Moreno, Luis</td>
<td>1099</td>
</tr>
<tr>
<td>Mitsakis, Evangelos</td>
<td>514</td>
</tr>
<tr>
<td>Moccaldi, Martina</td>
<td>1013</td>
</tr>
<tr>
<td>Moenius, Johannes</td>
<td>597</td>
</tr>
<tr>
<td>Moore, Adam</td>
<td>553</td>
</tr>
<tr>
<td>Morandini, Luca</td>
<td>756</td>
</tr>
<tr>
<td>Negi, Rohit</td>
<td>962</td>
</tr>
<tr>
<td>Neuhausler, Rosanna</td>
<td>313</td>
</tr>
<tr>
<td>Nguyen, Mai</td>
<td>896</td>
</tr>
<tr>
<td>Nino-Ruiz, Marcos</td>
<td>756</td>
</tr>
<tr>
<td>Northrop, William</td>
<td>360</td>
</tr>
<tr>
<td>O’Brien, Oliver</td>
<td>712</td>
</tr>
<tr>
<td>O’Sullivan, Aidan</td>
<td>276</td>
</tr>
<tr>
<td>Offenhuber, Dietmar</td>
<td>158</td>
</tr>
<tr>
<td>Owen, Andrew</td>
<td>1209</td>
</tr>
<tr>
<td>Park, Tae Hong</td>
<td>877</td>
</tr>
<tr>
<td>Pereira, Francisco</td>
<td>276</td>
</tr>
<tr>
<td>Peters, Jonathan</td>
<td>941</td>
</tr>
<tr>
<td>Pettit, Christopher</td>
<td>756</td>
</tr>
<tr>
<td>Pless, Robert</td>
<td>638</td>
</tr>
<tr>
<td>Prabhu, Vinay</td>
<td>962</td>
</tr>
<tr>
<td>Name</td>
<td>Page</td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Radke, John D.</td>
<td>313</td>
</tr>
<tr>
<td>Randolph, Bill</td>
<td>756</td>
</tr>
<tr>
<td>Rivera, Samuel</td>
<td>846</td>
</tr>
<tr>
<td>Rodrigues, Filipe</td>
<td>276</td>
</tr>
<tr>
<td>Rodrigues, Miguel</td>
<td>962</td>
</tr>
<tr>
<td>Rofe, Yodan</td>
<td>1188</td>
</tr>
<tr>
<td>Roth, Dan</td>
<td>846</td>
</tr>
<tr>
<td>Salanova Grau, Josep Maria</td>
<td>514</td>
</tr>
<tr>
<td>Saldarriaga, Juan Francisco</td>
<td>941</td>
</tr>
<tr>
<td>Saunier, Nicolas</td>
<td>1099</td>
</tr>
<tr>
<td>Schmidt-Poolman, Martine</td>
<td>313</td>
</tr>
<tr>
<td>Schnebele, Emily</td>
<td>1013</td>
</tr>
<tr>
<td>Shao, Zhenfeng</td>
<td>699</td>
</tr>
<tr>
<td>Shekhar, Shashi</td>
<td>360</td>
</tr>
<tr>
<td>Shen, Guoqiang</td>
<td>1075</td>
</tr>
<tr>
<td>Shroff, Ravi</td>
<td>608</td>
</tr>
<tr>
<td>Sicignano, Rosa</td>
<td>1013</td>
</tr>
<tr>
<td>Sinnott, Richard</td>
<td>756</td>
</tr>
<tr>
<td>Smith, Tom</td>
<td>201</td>
</tr>
<tr>
<td>Spielman, Seth</td>
<td>49, 184</td>
</tr>
<tr>
<td>St-Aubin, Paul</td>
<td>1099</td>
</tr>
<tr>
<td>Stamos, Iraklis</td>
<td>514</td>
</tr>
<tr>
<td>Steer, Celine</td>
<td>734</td>
</tr>
<tr>
<td>Stimson, Robert</td>
<td>756</td>
</tr>
<tr>
<td>Sui, Daniel</td>
<td>253</td>
</tr>
<tr>
<td>Sun, Shipeng</td>
<td>603</td>
</tr>
<tr>
<td>Tang, Zhenghong</td>
<td>33</td>
</tr>
<tr>
<td>Tasse, Dan</td>
<td>64</td>
</tr>
<tr>
<td>Thakuriah, Piyushimita</td>
<td>1, 4, 398</td>
</tr>
<tr>
<td>Thepvojojanapong, Niwat</td>
<td>106</td>
</tr>
<tr>
<td>Tilahun, Nebiyou</td>
<td>1, 4</td>
</tr>
<tr>
<td>Tobe, Yoshito</td>
<td>106</td>
</tr>
<tr>
<td>Tomko, Martin</td>
<td>756</td>
</tr>
<tr>
<td>Tufte, Kristin</td>
<td>808</td>
</tr>
<tr>
<td>Ukkusuri, Satish</td>
<td>1129</td>
</tr>
<tr>
<td>Usami, Itaru</td>
<td>106</td>
</tr>
<tr>
<td>Wan, Dan</td>
<td>1229</td>
</tr>
<tr>
<td>Wang, Haozhe</td>
<td>608</td>
</tr>
<tr>
<td>Wang, Shaowen</td>
<td>123</td>
</tr>
<tr>
<td>Wang, Yaoli</td>
<td>414</td>
</tr>
<tr>
<td>Wang, Zun-Jing</td>
<td>859</td>
</tr>
<tr>
<td>Waters, Nigel</td>
<td>1013</td>
</tr>
<tr>
<td>Widjaja, Ivo</td>
<td>756</td>
</tr>
<tr>
<td>Wood, Brittany</td>
<td>213</td>
</tr>
</tbody>
</table>
Work, Daniel 846
Wu, Jeremy 435
Wu, Laiyun 1258
Wu, Liang 1258

Xie, Xiao-Feng 859

Yan, Jacob 80
Yan, Wai Yeung 463
Yang, Ci 1149
Yazici, Anil 1229
Yin, Junjun 123
Yin, Li 699, 1258
Yu, Hongfeng 33

Zellner, Moira 1, 4
Zhan, Xianyuan 1129
Zhang, Rui 435
Zhang, Zhanmin 498
Zhao, Bo 253
Zhou, Xuesong 332
Zhou, Yanfu 33
## Keyword Index

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Page(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3Di Hydrodynamic Model</td>
<td>83</td>
</tr>
<tr>
<td>Accessibility</td>
<td>354, 1209</td>
</tr>
<tr>
<td>Acoustic Ecology</td>
<td>241</td>
</tr>
<tr>
<td>Activity-Based Models</td>
<td>153</td>
</tr>
<tr>
<td>Activity-Travel Pattern</td>
<td>101</td>
</tr>
<tr>
<td>Agent Based Simulation</td>
<td>263</td>
</tr>
<tr>
<td>Agglomeration</td>
<td>377</td>
</tr>
<tr>
<td>Air Pollution</td>
<td>173</td>
</tr>
<tr>
<td>American Community Survey</td>
<td>41</td>
</tr>
<tr>
<td>Analytics</td>
<td>102</td>
</tr>
<tr>
<td>API Data</td>
<td>155</td>
</tr>
<tr>
<td>App</td>
<td>181</td>
</tr>
<tr>
<td>Asymmetric Mapping</td>
<td>177</td>
</tr>
<tr>
<td>Automated Passenger Count</td>
<td>207</td>
</tr>
<tr>
<td>Automated Vehicle Location</td>
<td>207</td>
</tr>
<tr>
<td>Bayesian Networks</td>
<td>328</td>
</tr>
<tr>
<td>Big Data</td>
<td>1, 4, 201, 213</td>
</tr>
<tr>
<td></td>
<td>253, 294, 379, 435</td>
</tr>
<tr>
<td></td>
<td>485, 514, 528, 603</td>
</tr>
<tr>
<td></td>
<td>734, 780, 808, 877</td>
</tr>
<tr>
<td></td>
<td>896, 1129, 1308, 1327</td>
</tr>
<tr>
<td>Big Data Analytics</td>
<td>206</td>
</tr>
<tr>
<td>Bikeshare</td>
<td>64, 941</td>
</tr>
<tr>
<td>Building Energy</td>
<td>379</td>
</tr>
<tr>
<td>Built Environment</td>
<td>179</td>
</tr>
<tr>
<td>Carsharing Rental Choice</td>
<td>155</td>
</tr>
<tr>
<td>Categorization</td>
<td>40</td>
</tr>
<tr>
<td>Cellular Network Data</td>
<td>101</td>
</tr>
<tr>
<td>Census</td>
<td>41</td>
</tr>
<tr>
<td>Census Data</td>
<td>262</td>
</tr>
<tr>
<td>Centralized Data Platform</td>
<td>205</td>
</tr>
<tr>
<td>Challenges</td>
<td>4</td>
</tr>
<tr>
<td>China</td>
<td>125</td>
</tr>
<tr>
<td>Cities</td>
<td>4</td>
</tr>
<tr>
<td>Citizen Feedback Systems</td>
<td>40</td>
</tr>
<tr>
<td>Citizen Science</td>
<td>33</td>
</tr>
<tr>
<td>Civic Hacking</td>
<td>123</td>
</tr>
<tr>
<td>Cloud Computing</td>
<td>301</td>
</tr>
<tr>
<td>Clustering</td>
<td>379</td>
</tr>
<tr>
<td>Collective Intelligence</td>
<td>122</td>
</tr>
<tr>
<td>Collision Prediction</td>
<td>327</td>
</tr>
</tbody>
</table>
Community Detection 124, 1173
Comparisons 42
Complexity 100
Computer Vision 327
Consumer Popularity 378
Count Regression Model 329
Crowdsensing 37
Crowdsourcing 36, 379, 638
CyberGIS 38

Dashboard 204
Data Analysis 177
Data Analytics 209
Data Disaggregation 263
Data Fusion 41
Data Governance 40
Data Mining 180
Data Portal 204
Data Quality 174
Data Store 204
Data Validation 34
Data Visualization 35
Database 241, 1209
DBSCAN 180
Demographics 177
Design Implications 40
Digital Divide 260
Digital Infomediaries 123
Distance Decay Coefficient 124
Driving Behavior Analysis 172
Duration Model 155

Education 102, 608
Electric Vehicle 154
Emergency Evacuation Framework 301
Energy Policy 380
Environment 37
Environmental 261
Environmental Big Data 33
Environmental Sensing 241
Equity 260, 941, 1188
Evacuation Management 278

Fares 356
Feature Selection 379
Floating Car Data 171
Foursquare 378
Free-Floating Carsharing 155
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freight Transportation</td>
<td>326</td>
</tr>
<tr>
<td>Gauss-Markov Random Fields</td>
<td>262</td>
</tr>
<tr>
<td>Geospatial Analysis</td>
<td>177, 1013</td>
</tr>
<tr>
<td>GIS</td>
<td>36, 463, 1188</td>
</tr>
<tr>
<td>Global-Local</td>
<td>326</td>
</tr>
<tr>
<td>Google Street View Images</td>
<td>203</td>
</tr>
<tr>
<td>Governance</td>
<td>122</td>
</tr>
<tr>
<td>GPS</td>
<td>180</td>
</tr>
<tr>
<td>GPS Trajectories</td>
<td>43</td>
</tr>
<tr>
<td>GPS Trajectory</td>
<td>172</td>
</tr>
<tr>
<td>Graph Theory</td>
<td>354</td>
</tr>
<tr>
<td>Hackathon</td>
<td>181</td>
</tr>
<tr>
<td>Hedonic Regressions</td>
<td>175</td>
</tr>
<tr>
<td>Heteroscedasticity</td>
<td>46</td>
</tr>
<tr>
<td>Hidden Markov Model</td>
<td>101</td>
</tr>
<tr>
<td>High Resolution Data</td>
<td>278</td>
</tr>
<tr>
<td>High Spatial Resolution</td>
<td>354</td>
</tr>
<tr>
<td>Hotel</td>
<td>377</td>
</tr>
<tr>
<td>HTML5</td>
<td>301</td>
</tr>
<tr>
<td>Human Mobility</td>
<td>223</td>
</tr>
<tr>
<td>Industrial Systems</td>
<td>174</td>
</tr>
<tr>
<td>Informatics</td>
<td>210</td>
</tr>
<tr>
<td>Infrastructure Assessment</td>
<td>264</td>
</tr>
<tr>
<td>Innovations</td>
<td>4</td>
</tr>
<tr>
<td>Interactions</td>
<td>327</td>
</tr>
<tr>
<td>International Trade</td>
<td>326</td>
</tr>
<tr>
<td>Internet Data</td>
<td>377</td>
</tr>
<tr>
<td>Intra-Urban Migration</td>
<td>176</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Justice</td>
<td>261</td>
</tr>
<tr>
<td>Knowledge Discovery</td>
<td>4, 846</td>
</tr>
<tr>
<td>Land Use</td>
<td>155</td>
</tr>
<tr>
<td>Land Use Planning</td>
<td>100</td>
</tr>
<tr>
<td>LiDAR</td>
<td>83</td>
</tr>
<tr>
<td>Liveability</td>
<td>209</td>
</tr>
<tr>
<td>Local Government</td>
<td>260</td>
</tr>
<tr>
<td>Location Sensing Applications</td>
<td>44</td>
</tr>
<tr>
<td>Location Spoofing</td>
<td>45</td>
</tr>
<tr>
<td>Location-Based Services</td>
<td>223</td>
</tr>
<tr>
<td>Locational Analysis</td>
<td>174</td>
</tr>
<tr>
<td>Machine Learning</td>
<td>46, 877</td>
</tr>
<tr>
<td>Term</td>
<td>Page</td>
</tr>
<tr>
<td>-------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Mapping</td>
<td>83</td>
</tr>
<tr>
<td>Markets</td>
<td>261</td>
</tr>
<tr>
<td>Meeting Minutes</td>
<td>39</td>
</tr>
<tr>
<td>Mobile</td>
<td>38</td>
</tr>
<tr>
<td>Mobile Phone Data</td>
<td>124</td>
</tr>
<tr>
<td>Mobile Phones</td>
<td>37</td>
</tr>
<tr>
<td>Mobility Data</td>
<td>171</td>
</tr>
<tr>
<td>Mode Choice</td>
<td>356</td>
</tr>
<tr>
<td>Model</td>
<td>326</td>
</tr>
<tr>
<td>Modeling</td>
<td>177</td>
</tr>
<tr>
<td>Motion Pattern Learning</td>
<td>327</td>
</tr>
<tr>
<td>Neighborhood Data</td>
<td>41</td>
</tr>
<tr>
<td>Network Analysis</td>
<td>124</td>
</tr>
<tr>
<td>Node.js</td>
<td>301</td>
</tr>
<tr>
<td>Noise Mapping</td>
<td>38</td>
</tr>
<tr>
<td>Non-Linear Modeling</td>
<td>380</td>
</tr>
<tr>
<td>Non-Parametric Methods</td>
<td>209</td>
</tr>
<tr>
<td>OD Flow Matrix</td>
<td>353</td>
</tr>
<tr>
<td>Online Participatory Tool</td>
<td>64</td>
</tr>
<tr>
<td>Online Workbench</td>
<td>206</td>
</tr>
<tr>
<td>Open Data</td>
<td>123, 608, 712, 962</td>
</tr>
<tr>
<td>Open Governance</td>
<td>181</td>
</tr>
<tr>
<td>Open Government Data</td>
<td>278</td>
</tr>
<tr>
<td>Parallel Programming</td>
<td>208</td>
</tr>
<tr>
<td>Parametric Planning</td>
<td>100</td>
</tr>
<tr>
<td>Parcel Data</td>
<td>176</td>
</tr>
<tr>
<td>Participation</td>
<td>181</td>
</tr>
<tr>
<td>Participatory Sensing</td>
<td>40, 859</td>
</tr>
<tr>
<td>Pay-As-You-Drive-And-You-Save (Paydays)</td>
<td>172</td>
</tr>
<tr>
<td>Pedestrian</td>
<td>203</td>
</tr>
<tr>
<td>Performance Measures</td>
<td>173</td>
</tr>
<tr>
<td>Performance-Based Planning</td>
<td>207</td>
</tr>
<tr>
<td>Personalized Driving Data</td>
<td>172</td>
</tr>
<tr>
<td>Physical Activity</td>
<td>178, 638</td>
</tr>
<tr>
<td>Plan-Making</td>
<td>64</td>
</tr>
<tr>
<td>Planning</td>
<td>39</td>
</tr>
<tr>
<td>Planning Process</td>
<td>380</td>
</tr>
<tr>
<td>Plug-In Electric Vehicle Charging Infrastructure</td>
<td>101</td>
</tr>
<tr>
<td>Point Pattern Analysis</td>
<td>34</td>
</tr>
<tr>
<td>Pre-Kindergarten</td>
<td>177</td>
</tr>
<tr>
<td>Privacy</td>
<td>44</td>
</tr>
<tr>
<td>Private Data</td>
<td>174</td>
</tr>
<tr>
<td>Probe Data</td>
<td>171</td>
</tr>
<tr>
<td>Public Policy</td>
<td>177</td>
</tr>
<tr>
<td>Public Transit</td>
<td>354, 1209</td>
</tr>
<tr>
<td>Topic</td>
<td>Page</td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Synthetic Data</td>
<td></td>
</tr>
<tr>
<td>Taxi Data</td>
<td>328</td>
</tr>
<tr>
<td>Taxi Demand Modeling</td>
<td>329</td>
</tr>
<tr>
<td>Taxi GPS Data</td>
<td>329</td>
</tr>
<tr>
<td>Taxi Trip Generation</td>
<td>356</td>
</tr>
<tr>
<td>Taxicab</td>
<td>261</td>
</tr>
<tr>
<td>Text Mining</td>
<td>210</td>
</tr>
<tr>
<td>The Netherlands</td>
<td>378</td>
</tr>
<tr>
<td>Time Geography</td>
<td>43, 463</td>
</tr>
<tr>
<td>Traffic Signals</td>
<td>173</td>
</tr>
<tr>
<td>Transaction Data</td>
<td>175</td>
</tr>
<tr>
<td>Transit</td>
<td>207, 1209</td>
</tr>
<tr>
<td>Transit Accessibility</td>
<td>329</td>
</tr>
<tr>
<td>Transit Operations</td>
<td>173</td>
</tr>
<tr>
<td>Transportation</td>
<td>153</td>
</tr>
<tr>
<td>Transportation Data</td>
<td>208</td>
</tr>
<tr>
<td>Travel Behavior</td>
<td>178</td>
</tr>
<tr>
<td>Travel Time</td>
<td>356</td>
</tr>
<tr>
<td>Travel Time Estimation</td>
<td>328</td>
</tr>
<tr>
<td>Trip Detection</td>
<td>180</td>
</tr>
<tr>
<td>Twitter</td>
<td>39, 253</td>
</tr>
<tr>
<td>Uncertainty</td>
<td>41, 253</td>
</tr>
<tr>
<td>Urban Amenities</td>
<td>378</td>
</tr>
<tr>
<td>Urban Analytics</td>
<td>35, 1129</td>
</tr>
<tr>
<td>Urban Big Data</td>
<td>123, 1173</td>
</tr>
<tr>
<td>Urban Data</td>
<td>123</td>
</tr>
<tr>
<td>Urban Environmental Management</td>
<td>33</td>
</tr>
<tr>
<td>Urban Informatics</td>
<td>1, 4, 756</td>
</tr>
<tr>
<td>Urban Metrics</td>
<td>204</td>
</tr>
<tr>
<td>Urban Planning</td>
<td>4, 345</td>
</tr>
<tr>
<td>Urban Sentiments</td>
<td>39</td>
</tr>
<tr>
<td>Urban Traffic Control</td>
<td>223</td>
</tr>
<tr>
<td>Urban Transportation Systems</td>
<td>278</td>
</tr>
<tr>
<td>Urbanization</td>
<td>125</td>
</tr>
<tr>
<td>Usage-Based Insurance</td>
<td>172</td>
</tr>
<tr>
<td>User-Generated Data</td>
<td>34</td>
</tr>
<tr>
<td>Value At Risk</td>
<td>175</td>
</tr>
<tr>
<td>Vehicle Electrification</td>
<td>154</td>
</tr>
<tr>
<td>Visual Preference Survey</td>
<td>36</td>
</tr>
<tr>
<td>Visualization</td>
<td>102, 608</td>
</tr>
<tr>
<td>Viterbi Algorithm</td>
<td>101</td>
</tr>
<tr>
<td>Volunteered Geographic Information</td>
<td>38, 294</td>
</tr>
<tr>
<td>Walkability</td>
<td>203</td>
</tr>
<tr>
<td>Term</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>Webcams</td>
<td>179</td>
</tr>
<tr>
<td>WebGL</td>
<td>301</td>
</tr>
<tr>
<td>Yelp</td>
<td>34</td>
</tr>
</tbody>
</table>