Data and expertise
for improving cities

www.ubdc.ac.uk

Using new forms of data to analyse cycling activity

Transcript from webinar video recording

1

00:00:04,720 --> 00:00:11,120

[Muir Houston] So, let me introduce everyone to this session on using new forms of data to

2

00:00:11,120 --> 00:00:17,440

analyse cycling activity. Dr Jinhyun Hong is a Senior Lecturer in transportation planning

3

00:00:17,440 --> 00:00:24,720

in Urban Studies and leads the Transport and Infrastructure team at UBDC. Jin's research

4

00:00:24,720 --> 00:00:30,320

interests include interaction amongst the built environment, travel behaviour and air quality,

5

00:00:30,320 --> 00:00:38,480

transportation and planning, the built environment, safety and walking and travel survey techniques.

6

00:00:38,480 --> 00:00:45,280

I see we have participants from Australia, Austria, Belgium, China, Germany, Iraq, the Philippines, Russia,

7

00:00:45,280 --> 00:00:48,880

Turkey, Ukraine, and the UK. Sorry if I’ve missed any of you.

8

00:00:49,920 --> 00:00:55,120

As you will have seen this session is recorded and will be uploaded on the web in an accessible

9

00:00:55,120 --> 00:01:01,680

format at some point after the session. Details will be provided on the UBDC website.

10

00:01:01,680 --> 00:01:08,080

Also, check out the website for other resources, including how to access other data and

11

00:01:08,080 --> 00:01:14,240

other training and events delivered by the UBDC. As we've mentioned, cameras will be turned

12

00:01:14,240 --> 00:01:20,160

off and microphones muted to aid privacy and also for bandwidth reasons. Please use the Q&A

13

00:01:20,160 --> 00:01:26,240

facility to ask questions. These will be collated, and responses will be provided in the Q&A

14

00:01:26,240 --> 00:01:32,800

session. In terms of the session structure, Jin will give a presentation for around 30 minutes,

15

00:01:33,440 --> 00:01:39,920

which will be followed by a Q&A session of 10 to 15 minutes. We will then have a break between

16

00:01:39,920 --> 00:01:46,800

sessions for 15 minutes or so and then we will have our second session with a similar format -

17

00:01:46,800 --> 00:01:53,600

30 minutes presentation and again 10 to 15 minutes for questions and answers.

18

00:01:53,600 --> 00:01:59,760

So, I'd just like to introduce our presenter Jin Hong and I'll hand over to Jin.

19

00:02:02,400 --> 00:02:09,040

[Jin Hong] Thanks a lot Muir and thanks to all of you for joining today's webinar and I'm sorry for the

20

00:02:09,040 --> 00:02:16,800

delay. In this session, as Muir said, I will talk about crowdsourced cycling data,

21

00:02:16,800 --> 00:02:23,600

that is Strava data, and as a researcher how we have used the data for cycling studies.

22

00:02:24,560 --> 00:02:31,680

In the next session I will do some kind of tutorial, as you may know.

23

00:02:34,240 --> 00:02:45,200

Here's a brief background about the studies. So, the large benefits of

24

00:02:45,200 --> 00:02:51,600

cycling have been well documented. It could reduce the auto dependency, therefore reduce

25

00:02:51,600 --> 00:02:57,840

the level of congestion and emissions. It could also improve the public health because people are

26

00:02:57,840 --> 00:03:05,040

doing physical exercise while cycling. In addition, if you look at the travel surveys from different

27

00:03:05,040 --> 00:03:12,000

countries, you will notice that a substantial amount of the automobile trips are short trips.

28

00:03:12,960 --> 00:03:19,760

That means their travel distance is between two and five kilometres. What does

29

00:03:19,760 --> 00:03:26,800

that mean? This is quite a long distance for walking. However, this is a really reasonable travel

30

00:03:26,800 --> 00:03:34,400

distance for cycling. So, it implies that cycling can be a good alternative of the automobile.

31

00:03:35,360 --> 00:03:41,440

So, because of this huge benefit and the potential, many countries have used their substantial

32

00:03:41,440 --> 00:03:49,520

resources to improve the cycling environment and also increase cycling. And this is the same for

33

00:03:49,520 --> 00:03:56,880

the UK. The active travel - walking and cycling - is one of the priorities for the National Transport

34

00:03:56,880 --> 00:04:03,680

Strategies in the UK. In Scotland we have a really ambitious vision. Transport Scotland

35

00:04:03,680 --> 00:04:09,520

wants 10 percent of journeys to be made by bicycle by 2020. This is really ambitious

36

00:04:09,520 --> 00:04:16,000

because now we have one to three percent at most. And the cities are responsible for achieving this.

37

00:04:16,880 --> 00:04:27,440

So, again we are trying to promote cycling. In Glasgow, the local government have also introduced several

38

00:04:27,440 --> 00:04:35,760

measures and interventions to promote cycling. For example, they want to use the 2014 Commonwealth Games -

39

00:04:35,760 --> 00:04:42,320

this is the international sports games - as a catalyst to promote cycling,

40

00:04:42,320 --> 00:04:48,240

to increase the cycling. So, they provided several cycling infrastructure lanes

41

00:04:48,880 --> 00:04:56,080

before, during and after the Commonwealth Games. In addition, they also provide bike share programme Nextbike

42

00:04:56,080 --> 00:05:04,320

as you can see in the pictures. They are quite popular at this moment. However, as a planner if you

43

00:05:04,320 --> 00:05:11,040

want to make better cycling plans, you really need to understand the cycling patterns, cycling behaviour,

44

00:05:11,040 --> 00:05:18,480

and also, you need to understand how to use the proper data to evaluate the effectiveness of

45

00:05:18,480 --> 00:05:26,960

interventions. Unfortunately, these are very difficult because we don't have proper

46

00:05:26,960 --> 00:05:35,440

data. We have travel surveys. So, many people actually use travel surveys to examine travel behaviour.

47

00:05:36,000 --> 00:05:44,480

However, there are only a small number of cyclists in most cities and although its representative sample

48

00:05:44,480 --> 00:05:49,680

does travel surveys, they only include a small portion of the people from the population,

49

00:05:49,680 --> 00:05:57,680

compared to the population. What happens is, at the end, you may end up with 40 or 50 people who cycled

50

00:05:57,680 --> 00:06:05,280

in your travel survey for a metropolitan area. Then that's too small. You cannot really use that

51

00:06:05,280 --> 00:06:12,720

data to build some model or to analyse detailed cycling activities. We also have a manual and

52

00:06:12,720 --> 00:06:19,600

automatic count. So, for example, in Glasgow every year for two days they manually count how many

53

00:06:19,600 --> 00:06:26,960

cycles in and from the city centre and in some cities they installed the automatic counters

54

00:06:26,960 --> 00:06:34,640

to continuously measure the cycle activities. Again, these are very expensive hence infrequent. There

55

00:06:34,640 --> 00:06:41,440

are only a small number of automatic counters in the city because again it's very expensive.

56

00:06:41,440 --> 00:06:48,000

So, it's a really good ground to this data they are, but there are significant

57

00:06:48,000 --> 00:06:54,800

limitations if we want to use this data to examine the cycling patterns, cycling behaviour.

58

00:06:56,880 --> 00:07:04,320

Due to the technology improvements, now we have new forms of data and these data provide

59

00:07:04,320 --> 00:07:10,720

detailed cycling activities at the fine spatial and temporal scale. The Strava

60

00:07:10,720 --> 00:07:17,200

cycling app is one of them and I think it's one of the most popular cycling apps in the world.

61

00:07:17,840 --> 00:07:26,160

And they use GPS to track cyclist's journeys, so they know exactly what time and where the

62

00:07:26,160 --> 00:07:34,320

the Strava users are using cycling. So that is a really amazing data set and as time passes

63

00:07:34,320 --> 00:07:39,920

more people are using this app because it became popular. So, then what does that mean? The quality

64

00:07:39,920 --> 00:07:45,360

of the data will improve because there are more data.

65

00:07:47,200 --> 00:07:53,520

In addition, the data are already being collected all over the world because everyone can

66

00:07:53,520 --> 00:08:01,440

download the app. So, we can compare the same policy in different countries using

67

00:08:01,440 --> 00:08:07,280

the same data format and we also can use the same methodological approaches

68

00:08:07,280 --> 00:08:14,720

to different cases because the data structures are exactly the same. However, as a researcher we

69

00:08:14,720 --> 00:08:21,520

really need to understand then what could be the potential weaknesses of these emerging

70

00:08:21,520 --> 00:08:30,480

forms of data, which could influence our study. So, what are the weaknesses? The first thing

71

00:08:30,480 --> 00:08:38,640

is representativeness. I guess you already know this one. For example, in the Strava case,

72

00:08:38,640 --> 00:08:44,960

it's more likely young male people are more likely to use the Strava apps and they are more

73

00:08:44,960 --> 00:08:52,320

experienced cyclists than the normal cyclists. So, if their cycling patterns are different from those

74

00:08:52,320 --> 00:08:59,760

of the normal cyclists then the research, the analysis, could be biased. So, we could

75

00:08:59,760 --> 00:09:06,880

get incorrect conclusions. So, there are many people recently who have tried to employ

76

00:09:06,880 --> 00:09:14,800

the advanced analytics methodology methods to correct the bias. There are also special variations

77

00:09:15,760 --> 00:09:22,640

As I said, most cities have only a small number of cyclists and among them only a small

78

00:09:22,640 --> 00:09:31,040

number of people actually uses Strava apps. So, if we look at the popular roads

79

00:09:31,040 --> 00:09:36,880

you may find some people who use the Strava apps, but if you look at the less popular roads

80

00:09:37,840 --> 00:09:43,280

although there are actual cyclists you may not have any Strava users. So,

81

00:09:44,080 --> 00:09:51,200

it depends on where the place is. Some people argue that Strava data is more useful

82

00:09:51,200 --> 00:09:57,600

for urban areas where the level of cycling activities is high compared to the rural areas.

83

00:09:59,120 --> 00:10:04,640

Lack of social demographic information, mainly because of privacy issues. We all know that the

84

00:10:04,640 --> 00:10:10,240

social demographic factors are very important determinants of the travel behaviour like

85

00:10:10,240 --> 00:10:18,320

age, gender, income, education level and so on. However, we don't have that information.

86

00:10:18,320 --> 00:10:22,720

The last one that I want to talk about is regulation from the company.

87

00:10:22,720 --> 00:10:29,600

This is actually very important, and I will talk about the Strava case later. Because the data

88

00:10:29,600 --> 00:10:36,480

is owned by the private company, if they somehow change their regulation or their products because

89

00:10:36,480 --> 00:10:46,080

of privacy or to protect their app users, this is beyond our control. We cannot really request

90

00:10:46,080 --> 00:10:53,040

against their decision. So, we just need to accept it

91

00:10:53,040 --> 00:11:00,720

and they could influence our whole study or the methodologies that we have used before. And

92

00:11:00,720 --> 00:11:06,480

this can be another problem in the near future because now a lot of private companies

93

00:11:06,480 --> 00:11:14,320

collect their own data. There are also others. Then what are the chances.

94

00:11:14,320 --> 00:11:20,880

I already mentioned several things that is most up-to-date information right now they are also

95

00:11:20,880 --> 00:11:26,640

collecting the data, because several people are using the Strava apps it's cost saving they

96

00:11:26,640 --> 00:11:32,400

don't really pay anything and it's very detailed cycling activity at the fine spatial and temporal

97

00:11:32,400 --> 00:11:38,480

scales and again a growing number of users. So, that could improve the quality of the data.

98

00:11:39,920 --> 00:11:45,360

So, this was the brief introduction about the cycling studies and also Strava.

99

00:11:46,240 --> 00:11:54,480

Here, I want to introduce three published papers that use the Strava data from our side.

100

00:11:54,480 --> 00:12:01,680

And I hope this will give you some kind of idea about how we use the data. And these are

101

00:12:01,680 --> 00:12:09,040

the three research questions that we aim to answer. The first one is can crowdsourced cycling data

102

00:12:09,040 --> 00:12:15,920

be utilised for cycling behaviour studies? This is about the quality of the data, whether the

103

00:12:15,920 --> 00:12:25,200

Strava data are good enough for studying the cycling activities. The second one is, if yes then

104

00:12:25,200 --> 00:12:32,560

where commuting cyclists travel and what are the influential factors for their route choice? Because

105

00:12:32,560 --> 00:12:39,120

we have such detail on cycling activities in the whole city area.

106

00:12:39,760 --> 00:12:46,720

The last one is, do the new cycle infrastructure investments in Glasgow produce effective impacts?

107

00:12:47,280 --> 00:12:53,840

So, these are the three research questions that I'm going to introduce. So, what data and variables?

108

00:12:53,840 --> 00:13:01,760

We used multi-years of Strava data, four years of Strava data 2013, 2014, 2015 and 2016.

109

00:13:01,760 --> 00:13:09,360

By the way, these data are available in UBDC. You can get the data based on your request.

110

00:13:10,720 --> 00:13:19,360

Data used to be provided as origin and destination with route information at the output

111

00:13:19,360 --> 00:13:27,280

area level. Output area is a UK census area. It's pretty small. So, what it means is we know for

112

00:13:27,280 --> 00:13:36,880

each trip, where the trip starts - output area - and what output area the person travelled and where is

113

00:13:36,880 --> 00:13:46,080

the destination - output area. This is pretty nice data. Second one is a more detailed one - minute by minute

114

00:13:46,080 --> 00:13:53,040

link count. Link count is at the road segment. We know

115

00:13:53,040 --> 00:14:00,160

how many people cycled on a particular time, minute by minute, this is really detailed

116

00:14:00,160 --> 00:14:08,320

data. We also have information about waiting times at junctions and then aggregated demographic

117

00:14:08,320 --> 00:14:16,880

information - for example, age and gender for your city. They just give you an aggregated summary.

118

00:14:19,280 --> 00:14:30,160

Now, they changed the product because of privacy issues. From 2018 Strava Metro the company has

119

00:14:30,160 --> 00:14:37,440

provided binned count data. So, what does that mean? They aggregated cycling counts in five

120

00:14:37,440 --> 00:14:44,240

count buckets. For example, if counts are less than three or equal to three it becomes zero.

121

00:14:45,360 --> 00:14:51,040

if counts are between four and seven it becomes five. What's the implication?

122

00:14:51,840 --> 00:14:57,520

As I said, there are a small number of cyclists and among them a small number of people are using

123

00:14:57,520 --> 00:15:04,320

the apps. So, what it means is if you look at the whole city it's a file for daily or hourly data.

124

00:15:04,320 --> 00:15:10,720

You will see a lot of measured roads will only have one to three or zero

125

00:15:11,840 --> 00:15:21,920

Strava users. So, depending on the cities, one Strava user could represent 25 to 100 actual cyclists.

126

00:15:21,920 --> 00:15:27,440

And because they are binning this data, they just lose huge information.

127

00:15:28,240 --> 00:15:33,120

So, that could influence your level of aggregation, your analysis unit.

128

00:15:34,720 --> 00:15:40,320

Now, they only provide hourly aggregation in the lowest level of aggregation,

129

00:15:40,320 --> 00:15:47,520

not a minute by minute. I think it's because of that issue. So, that is a big issue. Again,

130

00:15:47,520 --> 00:15:53,840

if that happens again, the whole methodologies we have used before may not be available because

131

00:15:54,480 --> 00:15:59,760

as some methods use the more detailed temporal scale.

132

00:16:01,920 --> 00:16:08,160

So, we used four years of the Strava data but at the time the unbinned data

133

00:16:08,160 --> 00:16:13,280

was available so we used unbinned data for our own study.

134

00:16:15,520 --> 00:16:22,080

We also used a manual count of cyclists from cordon count carried out in Glasgow in the same time

135

00:16:22,080 --> 00:16:30,560

period. So, there are 38 locations and two days in general in September or per year. I will show you

136

00:16:30,560 --> 00:16:36,720

the location later. And we also use Glasgow cycle infrastructure data as you can see on the map that

137

00:16:36,720 --> 00:16:45,680

is the current infrastructure map. And then this is a river, and we see there are really good

138

00:16:45,680 --> 00:16:55,040

nice cycling infrastructure alongside the river and here and other parts. So, we can see some

139

00:16:55,040 --> 00:17:01,120

areas like this one - east side of Glasgow - where it's the most deprived area, there are

140

00:17:01,120 --> 00:17:06,720

not many good cycling infrastructure. So, we can see some kind of special inequality issues here.

141

00:17:09,440 --> 00:17:14,240

So, for the research question one, about the quality of the data, how can we check

142

00:17:14,240 --> 00:17:21,440

the quality? We can check the ground truth data, which is a manual count data, with Strava data

143

00:17:21,440 --> 00:17:29,280

So, in the map you can see the 38 locations and that's the locations where people check the

144

00:17:29,280 --> 00:17:37,200

number of cyclists. And we know the location and time of the number of actual cyclists

145

00:17:37,200 --> 00:17:44,400

and we know exactly time and location from the Strava data. So, we compare them.

146

00:17:45,200 --> 00:17:52,160

That's how we do that. So, there are two types of analysis. One is a correlation analysis and the

147

00:17:52,160 --> 00:17:58,240

other one is simple linear regression model. The equation shows the simple linear regression model.

148

00:17:58,240 --> 00:18:06,800

y represents the number of cyclists from the cordon count, so ground truth data.

149

00:18:06,800 --> 00:18:13,600

And x Strava means the number of Strava cycling trips with a simple linear regression model. And

150

00:18:13,600 --> 00:18:19,280

for this analysis we use time period or three time periods. What it means is we aggregate the

151

00:18:19,280 --> 00:18:26,640

count for am peak, afternoon and pm peak. The reason is, we found later if we aggregate

152

00:18:26,640 --> 00:18:34,720

that level we could have the data quality improve and be much better. So, the total

153

00:18:34,720 --> 00:18:42,560

sample size is 684 because we have 38 locations, three time periods and two days and three years.

154

00:18:44,400 --> 00:18:49,440

That is for the first research question. So, for the second research question

155

00:18:49,440 --> 00:18:54,960

we use the 2016 Strava data but OD matrix. We can construct OD matrix

156

00:18:54,960 --> 00:19:02,560

because we know the output area of the origin and destination for each trip. So, we compare

157

00:19:02,560 --> 00:19:09,840

the routes taken by commuting cyclists. That means the Strava data, raw Strava data,

158

00:19:10,560 --> 00:19:17,680

with the route they would take if they minimised their travel distance. So, how can we do that? We

159

00:19:17,680 --> 00:19:25,360

use the traffic assignment model and the estimate based on the same OD matrix - the shortest travel

160

00:19:25,360 --> 00:19:33,440

distance path. And then find out the traffic volume per each link edge and we compare them.

161

00:19:34,160 --> 00:19:41,520

And we also use Google Maps and local knowledge to figure out why some roads are popular why some

162

00:19:41,520 --> 00:19:49,840

roads are less popular. For research question three we use the four years of

163

00:19:49,840 --> 00:19:58,720

Strava data and we calculate the total number of Strava trips per output level per month.

164

00:19:58,720 --> 00:20:06,080

It's monthly total. Why? As I will show you later, as we increase the level

165

00:20:06,080 --> 00:20:12,480

of aggregation the data quality becomes much better. So, monthly or average we think is really

166

00:20:12,480 --> 00:20:21,680

good enough for this study. We used output area as analytical units. Then we use the fixed effect

167

00:20:21,680 --> 00:20:28,080

poisson panel regression model as you're shown in the equation. So, the dependent variable is

168

00:20:28,080 --> 00:20:36,400

number of cycling trips in area i, output area i in month t. This is the four years of data, so panel

169

00:20:36,400 --> 00:20:45,040

data. And we have four infrastructure, you can see on the map. And then this one is interesting

170

00:20:45,040 --> 00:20:52,400

because the longest one connects the suburban area to the city centre. Here is the city centre.

171

00:20:55,280 --> 00:21:00,560

So, this is the result for research question one - a correlation analysis and simple linear

172

00:21:00,560 --> 00:21:05,840

regression model. We have different levels of aggregation because we think if we

173

00:21:05,840 --> 00:21:11,520

aggregate more and more, the data become larger and larger, we will have more counts

174

00:21:11,520 --> 00:21:18,800

so we may have a better kind of result. So, the lowest level of aggregation is hourly

175

00:21:18,800 --> 00:21:23,760

and the last one is the two days because we only have two days of cordon count data.

176

00:21:24,800 --> 00:21:32,880

And correlation, even hourly, we find point seven eight almost 0.8. This is very high. It

177

00:21:32,880 --> 00:21:41,040

gives us a positive signal. For the two days of aggregation we have almost 0.9, which is a really

178

00:21:41,040 --> 00:21:48,400

good kind of indicator. When we look at the linear regression model result, the estimates

179

00:21:48,400 --> 00:21:53,920

for the Strava trips our independent variable is positive and very significant

180

00:21:55,120 --> 00:22:02,000

and the adjusted R square is 0.74. By using this one variable this simple linear regression model

181

00:22:02,000 --> 00:22:10,000

can explain the 74 percent of the variations in the cordon count data. That is a huge one.

182

00:22:10,640 --> 00:22:17,920

So, these two results gives us kind of part of a signal. Yes, Strava data could be used to examine

183

00:22:17,920 --> 00:22:24,480

the spatial variation of the cycling patterns. The full graph shows the relationship between Strava

184

00:22:24,480 --> 00:22:31,760

data and the cordon count. So, x axis is number of Strava data, y axis is number of

185

00:22:31,760 --> 00:22:38,160

cordon count. And you can see hourly aggregation there are a lot of noise at the bottom part left

186

00:22:38,160 --> 00:22:44,560

bottom because again there are only small number of people who cycle. But if we aggregate, we see

187

00:22:44,560 --> 00:22:50,080

more clear pattern of the linear regression and noise becomes smaller. So, based on this

188

00:22:50,080 --> 00:22:58,080

result we concluded that, yes, it could be good proper data for cycling studies.

189

00:22:59,680 --> 00:23:06,560

For the second research question, we compare the short routes and actual routes. The graph on your

190

00:23:06,560 --> 00:23:14,240

left panel it's actual Strava data, so that's where people really cycled. So, we see alongside

191

00:23:14,240 --> 00:23:20,880

the river there is some concentration, they are popular. That makes sense because amenity is one of

192

00:23:20,880 --> 00:23:28,320

the main factors for cycling and also safe cycling lane infrastructure. For the shortest

193

00:23:29,360 --> 00:23:35,600

paths you see more evenly distributed kind of patterns. And that also makes sense but

194

00:23:35,600 --> 00:23:42,800

it's already hard to see. So, to make a better vision we calculate the difference between them.

195

00:23:43,680 --> 00:23:51,920

Here. The red means it's more popular, so there are more cyclists,

196

00:23:51,920 --> 00:23:57,840

cycling trips happened on these roads compared to predicted ones. And the black means

197

00:23:57,840 --> 00:24:05,360

they are unpopular, less popular, and the thickness means the size of difference. So, we see

198

00:24:05,360 --> 00:24:11,760

alongside the river where there are good cycling infrastructure, there are really many people cycling.

199

00:24:12,640 --> 00:24:19,360

But around the city centre area there are some red lines and black lines, but they are

200

00:24:19,360 --> 00:24:24,880

very thin. It means their difference doesn't really get different. That totally makes sense.

201

00:24:26,160 --> 00:24:32,640

Right? So, then we see based on our local knowledge, we want to figure out why some roads are less

202

00:24:32,640 --> 00:24:42,080

popular. And here's one example this is one from the southside of the city centre and other

203

00:24:42,080 --> 00:24:47,440

roads are very popular but this road, although it's very straightforward, it's straight

204

00:24:47,440 --> 00:24:54,880

length, it's less popular. And see through the Google Map we want to see

205

00:24:54,880 --> 00:25:00,960

what's the problems. And these are the two pictures. We can see the bus stops, traffic lights with the

206

00:25:00,960 --> 00:25:05,588

pedestrian crossing and street parking. See, there is a lot of street parking and these cars

207

00:25:05,588 --> 00:25:11,021

there and no cycling infrastructure. Those are the factors often mentioned in previous

208

00:25:11,021 --> 00:25:18,080

studies, in particular studies, as a barrier of the cycling. So, we see yeah, maybe that's the reason.

209

00:25:18,080 --> 00:25:21,920

Another one, this is the east part of Glasgow

210

00:25:21,920 --> 00:25:29,200

and actually there is a cycling lane, however shared with buses and you can see a lot of cars parked.

211

00:25:29,760 --> 00:25:36,240

And also, the built environment. This area is one of the most deprived areas - high crime rate,

212

00:25:36,880 --> 00:25:43,840

the safety issue. Again, these are often mentioned as barriers for cycling and we can find those

213

00:25:43,840 --> 00:25:51,520

kinds of potential factors. So, when we do this analysis, we think this will be really good

214

00:25:51,520 --> 00:25:57,920

and simple tools for planners. They can really easily see where are the popular roads, where are

215

00:25:57,920 --> 00:26:02,800

the less popular roads and what are the potential reasons. They can understand their cities better.

216

00:26:04,880 --> 00:26:12,960

For the research question three, we use the model and these are the four investments. And we

217

00:26:12,960 --> 00:26:19,520

remove the time trend and then see how the total monthly count changed and the

218

00:26:20,080 --> 00:26:26,400

line here, you see the different location of the line, that's when the

219

00:26:26,400 --> 00:26:34,720

new infrastructure was open to the public. Right? And we see the three of them have

220

00:26:34,720 --> 00:26:42,080

a pretty positive increase after the new infrastructure were opened. However, this one

221

00:26:42,080 --> 00:26:47,840

the Routes to Cathkin 1, which actually connect the city centre to the suburban area, the longest

222

00:26:47,840 --> 00:26:56,960

one, has kind of a negative trend. So, after these basic stats we jump into the model and then

223

00:26:56,960 --> 00:27:04,320

these are the research from our model. We first measure the overall effect, and you can see it's

224

00:27:04,320 --> 00:27:12,240

there is a positive impact and it's about eight percent increase. You

225

00:27:12,240 --> 00:27:19,440

take the exponential to interpret the result. But the P-value is 0.08 which is greater than 0.05 so it's

226

00:27:19,440 --> 00:27:28,080

not really statistically significant. However, if you examine the separate effects, we notice that

227

00:27:29,200 --> 00:27:33,840

three among four have very positive impacts and they are very significant.

228

00:27:35,120 --> 00:27:41,840

It means a 12 percent to 18 percent increase after new infrastructure were introduced.

229

00:27:42,480 --> 00:27:48,960

and the right one, the first one, Routes to Cathkin 1 has a negative kind of impact

230

00:27:48,960 --> 00:27:54,320

compared to the other output area where there's no infrastructure. So, that is the

231

00:27:54,320 --> 00:28:00,880

reason why we didn't really get the significant result for the overall effect. And these three

232

00:28:00,880 --> 00:28:06,800

new infrastructure, they are close to the city centre area and they include the segregated

233

00:28:06,800 --> 00:28:12,480

lanes. So, they could provide some kind of policy implication. If you want to get a short-term impact

234

00:28:12,480 --> 00:28:18,880

then better to build more cycling lanes maybe near the city centre where the

235

00:28:20,000 --> 00:28:29,120

level of cycling activities is high and also the segregated lens could be important. So, this is

236

00:28:29,120 --> 00:28:35,440

the end of the slides. If you want to look more in detail about each of the studies and methodologies

237

00:28:35,440 --> 00:28:41,520

you can look at these three papers. They are open access, so you can download for free

238

00:28:41,520 --> 00:28:49,040

and this slide will be available for you, so you can get the full information. Thank you very much.

239

00:28:51,680 --> 00:28:58,000

[Muir Houston] Thank you very much Jin. We have some questions here. If I read them out one by one Jin, you

240

00:28:58,000 --> 00:29:05,040

could maybe try and answer them? So, the first question is, and I think you touched on this.

241

00:29:05,040 --> 00:29:12,000

How much Strava use the apps for commuting rather than the kind of lycra warriors that we see racing

242

00:29:12,000 --> 00:29:17,440

about the city that you mentioned, how many are kind of normal commuters if you like?

243

00:29:18,640 --> 00:29:25,760

[Jin Hong] So, in our case it was almost five percent of people are among the whole cyclists,

244

00:29:25,760 --> 00:29:33,120

five percent of the people are using the Strava apps and that's for our city

245

00:29:33,120 --> 00:29:40,960

but it depends on city by city. [Muir Houston] Ok, thanks for that. Now probably there might be some

246

00:29:40,960 --> 00:29:46,560

answer in the link you gave for your papers, is there in terms of cycling research case studies a journal?

247

00:29:47,440 --> 00:29:52,160

Is there transport planning or any other kind of common journals that you

248

00:29:52,160 --> 00:29:57,840

tend to publish in Jin? [Jin Hong] Yeah, so we published several papers in several leading transport

249

00:29:57,840 --> 00:30:05,280

journals. Transport Geography, Transportation and then Environment Planning A and B. So, we also

250

00:30:05,280 --> 00:30:11,200

published some paper in Transport Research part A, which is a top transport journal. So,

251

00:30:11,200 --> 00:30:17,840

you can find more information on my website or on the UBDC website. [Muir Houston] That's great thanks. And

252

00:30:18,640 --> 00:30:25,360

now the next one: have you tried to identify trips in cycling mode from other mobile phone

253

00:30:25,360 --> 00:30:34,480

data? For example, Telefonica or Voda UK instead of Strava. [Jin Hong] No, actually we haven't because

254

00:30:35,520 --> 00:30:40,960

now we are trying to get some mobile phone data, Urban Big Data Centre, we try to buy

255

00:30:40,960 --> 00:30:46,080

some mobile phone data, then we couldn't do it. That is actually the next step that we want to do

256

00:30:46,080 --> 00:30:51,920

because we want to also look at the other data sources. So, currently we are planning to buy

257

00:30:51,920 --> 00:31:00,800

some mobile phone data and we will also plan to conduct a survey with the apps

258

00:31:00,800 --> 00:31:06,160

so we can check the people's location and their trips. So, then we can use this

259

00:31:06,160 --> 00:31:14,560

data to detect mode choice from the mobile phone data. So, we hope we can have another session later

260

00:31:14,560 --> 00:31:21,040

using these new forms of data. [Muir Houston] Good stuff. Now another question here specifically about somebody

261

00:31:21,040 --> 00:31:26,160

who maybe knows Glasgow. In terms of popularity of routes for cycling, was consideration given

262

00:31:26,160 --> 00:31:31,680

to the physical and environmental factors e.g. topography or gradient? I think what they're

263

00:31:31,680 --> 00:31:37,760

meaning is, in Glasgow some of the streets are very big, long hills on them. Did people make a

264

00:31:37,760 --> 00:31:43,680

decision, did you see any of that Jin where people took a route because it was an easier cycle?

265

00:31:44,240 --> 00:31:49,920

[Jin Hong] Yes. Yes, so I will show you in the tutorial, actually. I will show you how

266

00:31:49,920 --> 00:31:56,000

to produce the maps using the Strava data and you can see the majority of the activity happens

267

00:31:56,000 --> 00:32:02,240

alongside the river which is very nice infrastructure it provides and very flat.

268

00:32:03,040 --> 00:32:10,240

And then you will see some kinds of less popular roads in some areas, very hilly areas.

269

00:32:11,200 --> 00:32:17,760

So, you can see, I mean you can easily examine the kinds of patterns by using Strava data

270

00:32:17,760 --> 00:32:25,200

for your own city. [Muir Houston] Good stuff. Now one, I'm not sure how much detail we provide here.

271

00:32:25,200 --> 00:32:31,920

May I ask how much Strava charged for the data and what is the attitude of people in Strava towards

272

00:32:31,920 --> 00:32:38,880

research projects or are they very commercial-focused? [Jin Hong] Yeah, this is the one that we may ask

273

00:32:38,880 --> 00:32:47,200

to Andrew who's the administration manager in UBDC, senior manager. I think

274

00:32:48,240 --> 00:32:57,520

we somehow spent 60k to buy the data for the Glasgow area for several years. But that depends

275

00:32:57,520 --> 00:33:04,960

on the license because, as you see, we purchase the data and so everyone who wants Strava data

276

00:33:05,520 --> 00:33:11,200

they can request and they can get it for free. So, if we are interested in the Strava data for

277

00:33:11,200 --> 00:33:18,560

Scotland especially, you can get the whole years of data from Urban Big Data Centre. And

278

00:33:18,560 --> 00:33:26,720

they are really eager to engage with academics. I want to say because we also have several

279

00:33:26,720 --> 00:33:32,640

conversations with them and they all know our studies. Sometimes they blog our studies.

280

00:33:33,600 --> 00:33:40,720

The problem is the current situation about safety. So, they try to change their data products

281

00:33:40,720 --> 00:33:46,240

and this is because its associated with running their companies.

282

00:33:46,800 --> 00:33:51,840

There could be some kind of issue. But in general, they are very friendly.

283

00:33:54,800 --> 00:34:01,200

[Muir Houston] Another question Jin, can you recommend some resources to study the traffic assignment model?

284

00:34:02,400 --> 00:34:07,280

[Jin Hong] There are many actually. There are books and there are papers, but you need to really learn

285

00:34:07,280 --> 00:34:13,360

the basic statistics model inside the traffic assignment model and also you need to

286

00:34:13,360 --> 00:34:19,600

run, if you want to only know the traffic assignment model you can just

287

00:34:19,600 --> 00:34:26,720

use, there are already code available online and also if you

288

00:34:26,720 --> 00:34:33,840

the easiest way is to use the travel demand models software

289

00:34:34,880 --> 00:34:42,080

TransCAD or VISUM. Then you can easily actually do that if you have the data. So yeah there are

290

00:34:42,080 --> 00:34:47,840

a lot of books if you can just google it or you can find several books or papers.

291

00:34:48,880 --> 00:34:58,560

[Muir Houston] And a question about socio-economic variables - do you use that much Jin? [Jin Hong] No,

292

00:34:58,560 --> 00:35:04,320

again, this is not actually included in the Strava data. They only provide, for example in the city of

293

00:35:04,320 --> 00:35:11,680

Glasgow, they only provide the distribution of the age and gender. So, how many females are in that

294

00:35:11,680 --> 00:35:19,040

area for that data. So, this is very aggregated data. So, we could actually use that data to

295

00:35:20,560 --> 00:35:26,000

check with the census data and see who are using the Strava apps. However, we cannot

296

00:35:26,000 --> 00:35:32,080

really use that information for the analysis because we don't really know for each trip.

297

00:35:34,480 --> 00:35:39,600

[Muir Houston] And one perhaps related to the current restrictions where there seems to have been a

298

00:35:39,600 --> 00:35:46,080

bit of an increase certainly in bike sales anyway. Do you think maybe in these

299

00:35:46,080 --> 00:35:53,440

circumstances it's easier to encourage people to cycle or walk more? [Jin Hong] Yeah, I think so and actually

300

00:35:53,440 --> 00:36:01,760

interestingly we have one draft

301

00:36:01,760 --> 00:36:10,080

paper that examines the cycling patterns after COVID-19 lockdown in the UK and you

302

00:36:10,080 --> 00:36:17,680

we saw a significant increase in terms of cycling activities. And yes, I think that's because

303

00:36:17,680 --> 00:36:24,960

of the current situation as well as the new situation, like for bike share programmes.

304

00:36:24,960 --> 00:36:31,520

They provide kind of free rides and also there are a lot of people who bought a bicycle. So yes, I

305

00:36:31,520 --> 00:36:39,280

think so. [Muir Houston] And the next one Jin, what proportion of all cyclists use Strava? Or to put it another

306

00:36:39,280 --> 00:36:45,840

way, how does your sample size for Strava data compare to the sample sizes for manual count data?

307

00:36:46,800 --> 00:36:51,440

[Jin Hong] That's so great. So again, as I said, I think that's the same for the first question.

308

00:36:51,440 --> 00:37:00,560

In our case it is about five percent, but again it varies city by city. So that's

309

00:37:00,560 --> 00:37:06,560

the reason why I said in one Strava user you could represent 25 people, actual cyclists, or

310

00:37:06,560 --> 00:37:12,640

sometimes, it depends on the city, it could be like 200 actual cyclists. And also, it depends on the

311

00:37:12,640 --> 00:37:17,840

location. It's urban area versus the rural area because, again, the spatial variation.

312

00:37:19,840 --> 00:37:27,440

[Muir Houston] And another one. Could average speed be used to differentiate between the sports cyclists

313

00:37:27,440 --> 00:37:34,160

and the commuters maybe? [Jin Hong] I think the easiest one is using the time. If commuters, there are certain

314

00:37:34,160 --> 00:37:40,560

times that they are using the cycles like am and pm peak. That's better I think in terms

315

00:37:40,560 --> 00:37:46,320

of separating the commuter trips and non-commuter trips and in Strava data they actually indicate

316

00:37:46,320 --> 00:37:51,840

whether they are commuting or not. So, if we are using the Strava data it's easy. If we are

317

00:37:51,840 --> 00:37:58,320

using other data sources, it's better to use the time, am peak and pm peak, to separate the commuting trips.

318

00:38:00,720 --> 00:38:06,480

[Muir Houston] Any methods for correcting the self-selection bias inherent in Strava

319

00:38:06,480 --> 00:38:13,280

data? [Jin Hong] So, there are currently several papers. They use some statistical models and use other

320

00:38:13,280 --> 00:38:19,840

built environments or other data sets and build some models to correct this bias. And there

321

00:38:19,840 --> 00:38:25,360

are also some studies, they are conducting their own survey and ask people whether they are

322

00:38:25,360 --> 00:38:32,320

using Strava apps or not and then their activities, patterns. And also, they

323

00:38:32,320 --> 00:38:40,480

measure the manual counts for different locations. And they use this whole

324

00:38:40,480 --> 00:38:47,760

kind of information together to try to correct the bias in the Strava data. So yes, there are papers

325

00:38:47,760 --> 00:38:55,920

it's not simple to understand, but there are people who are working on this issue. [Muir Houston] And here's an

326

00:38:55,920 --> 00:39:02,640

interesting one from somebody, my own research is in the area of crowdsourcing of qualitative data

327

00:39:02,640 --> 00:39:07,920

and experiences people have whilst traveling this would be a great addition to the quantitative data

328

00:39:07,920 --> 00:39:12,560

understanding the why as well as the what. Have you worked with anyone in adding that

329

00:39:12,560 --> 00:39:20,720

type of data through, for example, a bespoke app? [Jin Hong] Currently for the cycling, no. It was very hard to

330

00:39:22,320 --> 00:39:29,360

get the data based on our current situation. So, for the cycling we only have the Strava data.

331

00:39:29,360 --> 00:39:36,160

That was the main one, but again we are trying to purchase some mobile phone data

332

00:39:36,160 --> 00:39:43,360

and then if that is successful then we could actually use this new data to add new

333

00:39:43,360 --> 00:39:50,800

information. [Muir Houston] I think as well Jin, did Catherine's study not use ask people to keep a travel diary?

334

00:39:51,600 --> 00:39:57,360

[Jin Hong] That's the iMCD survey. That's a travel survey, household survey, so that's a little bit different.

335

00:39:57,360 --> 00:40:04,000

That's traditional survey data and not really the new forms of data like apps.

336

00:40:05,840 --> 00:40:11,120

But we do have another data set that we have at Urban Big Data Centre is

337

00:40:11,120 --> 00:40:17,520

iMCD survey which actually Muir just mentioned. There are survey, so it's a representative sample

338

00:40:17,520 --> 00:40:24,560

of the whole metropolitan area of Glasgow. And also, some 300 people carried a GPS

339

00:40:24,560 --> 00:40:31,040

life logging device for one week. So that's another data set we have. We have their

340

00:40:31,040 --> 00:40:38,640

travel diary and we have their GPS trajectory and also life logging picture data.

341

00:40:38,640 --> 00:40:44,480

So, if we may use those kinds of data together for the research

342

00:40:45,520 --> 00:40:52,320

to get more information. [Muir Houston] And actually Professor Lido, who works on that data, is giving one of

343

00:40:52,320 --> 00:40:56,880

the data dives, I think it's next week sometime. It'll be on the website, the details of that,

344

00:40:57,680 --> 00:41:09,360

if anybody's particularly interested in that. With regards to cyclist's safety Jin, on the routes, have you looked into how this could be understood potentially laying crowd-sourced data with STATS19

345

00:41:09,360 --> 00:41:16,480

accident information? [Jin Hong] We haven't done it, but we know there are studies and yes that's

346

00:41:16,480 --> 00:41:25,520

totally possible. We can also look at in Scotland and the UK SIMD, we know these

347

00:41:26,400 --> 00:41:32,560

crime rates of the areas so we can also see how they range between the crime rates and the

348

00:41:32,560 --> 00:41:37,040

cycling activities. And also look at the location, as the person asked.

349

00:41:39,040 --> 00:41:42,640

[Muir Houston] And just in terms of your work with Glasgow City Council Jin,

350

00:41:43,200 --> 00:41:49,120

are they quite receptive to use this data to help guide their cycling policies and plan

351

00:41:49,120 --> 00:41:55,360

infrastructure, for example new cycling lanes and stuff like that? [Jin Hong] They are very supportive all the

352

00:41:55,360 --> 00:42:02,160

infrastructure data they provide us for our own research and we also provide our research to them.

353

00:42:02,160 --> 00:42:07,680

And we are currently under discussion how we can help them to make a better cycling plan.

354

00:42:07,680 --> 00:42:13,280

And they are really amazing people, they are really kind and they

355

00:42:14,160 --> 00:42:18,800

are happy to collaborate with us.

356

00:42:22,240 --> 00:42:30,400

[Muir Houston] The next one, concern with using Strava data for making planning decisions as it could

357

00:42:30,400 --> 00:42:36,240

perpetuate transport inequalities by making women, older, lower income cyclists more invisible.

358

00:42:37,920 --> 00:42:46,320

[Jin Hong] Yes, so that is a reason why people try to correct the bias actually, by

359

00:42:46,320 --> 00:42:56,720

using other data sources. However, as we showed in the study, the general pattern is

360

00:42:56,720 --> 00:43:03,040

anyway, aggregated levels of cycling patterns. That could give us some kind of good policy implication

361

00:43:03,040 --> 00:43:11,200

and yes, I admit that, yes it could be the issue. [Muir Houston] Now I think I might know

362

00:43:11,200 --> 00:43:16,720

the answer to this one, can you track individuals over time? For example, can you see if an individual

363

00:43:16,720 --> 00:43:21,760

does the same journey every day of the week or is it just because of that binned data that's

364

00:43:22,320 --> 00:43:28,960

restricted or can you do it in the older data Jin? [Jin Hong] No actually it's not easy because we know

365

00:43:28,960 --> 00:43:35,280

for each trip the cycling routes, but that doesn't guarantee the same person has

366

00:43:35,280 --> 00:43:41,120

the same id, they don't have that kind of thing. Because, yes, if we can do that, that's a very serious

367

00:43:41,120 --> 00:43:48,000

privacy issue and they don't like it and we also don't like it. So no, I don't think that's easy,

368

00:43:48,000 --> 00:43:55,600

that's possible. And especially with the bin data, so I don't think that's possible.

369

00:43:57,280 --> 00:44:04,480

[Muir Houston] And a question here about, I know UBDC is doing some work on

370

00:44:04,480 --> 00:44:13,520

CCTV data for Glasgow City Council, could you use CCTV data rather than cordon counts as some kind

371

00:44:13,520 --> 00:44:20,000

of validation? [Jin Hong] Yes, that's what we are trying to do right now. In the CCTV project we have tried

372

00:44:20,000 --> 00:44:27,200

to identify the pedestrians and cyclists based on the CCTV data. If that is successful, I'm pretty

373

00:44:27,200 --> 00:44:34,240

sure we'll get it, then we can compare that data with our Strava data and then

374

00:44:34,240 --> 00:44:41,920

get more validation, you know further validation work. [Muir Houston] And what is the last

375

00:44:41,920 --> 00:44:48,720

question for this session I think, could you use Strava to perhaps plan the

376

00:44:48,720 --> 00:44:54,240

integration of the bicycle as transport in cities where the bicycle use is just emerging? So maybe a

377

00:44:54,240 --> 00:45:01,040

link back to this kind of covid restrictions and a lot more increase in people cycling in

378

00:45:01,040 --> 00:45:06,160

cities that don't have the infrastructure. Could Strava or something like that be helped to try...

379

00:45:08,080 --> 00:45:12,160

[Jin Hong] Yes, I think so because the first thing that you need to do is you need to understand the

380

00:45:12,160 --> 00:45:18,720

cycling patterns because there is reason why people use certain roads. But if you can know

381

00:45:18,720 --> 00:45:24,880

the whole picture at once, in a simple way, that will really help you to make a better plan.

382

00:45:24,880 --> 00:45:30,320

So yes, so that's another benefit of the Strava apps because that's available for everywhere

383

00:45:30,320 --> 00:45:36,080

in the world. So, you can, for example, there are some studies they're trying to find out,

384

00:45:36,080 --> 00:45:44,240

examine the cycling patterns in Africa where there's no real data and no infrastructure.

385

00:45:44,240 --> 00:45:50,320

But I think that's really one potential benefit of this kind of data. You

386

00:45:50,320 --> 00:45:56,560

can easily see the cycling patterns although there could be some bias, but you can easily see

387

00:45:56,560 --> 00:46:04,800

what roads are popular and why. [Muir Houston] That's great Jin, thank you very much. That's all of

388

00:46:04,800 --> 00:46:10,960

the questions for this session. Now I think we're going to have a break before the more

389

00:46:10,960 --> 00:46:20,800

practical session, so on my clock it's 10:55. If we say 11:15 we'll reconvene for the second

390

00:46:20,800 --> 00:46:28,000

session. So, thanks everyone for taking part. We'll return at 11:15 prompt for the next session and

391

00:46:28,000 --> 00:46:41,840

again, Jin will give a presentation and we can have another Q&A session. So, thanks just now. [Jin Hong] Thank you.

392

00:46:48,560 --> 00:46:53,200

[Muir Houston] Hi folks, welcome back to everyone. I hope you've all joined us again

393

00:46:53,920 --> 00:47:00,960

as we start this second session, which is a more hands-on practical and

394

00:47:01,520 --> 00:47:07,840

how to work through some of the examples that Jin talked about in the

395

00:47:07,840 --> 00:47:15,680

earlier session. So, same as before - questions and answers in the tab and we'll take them

396

00:47:16,320 --> 00:47:23,360

again at the end. So, I'll just hand over to Jin again and thanks for joining us. [Jin Hong] Thank you.

397

00:47:24,000 --> 00:47:29,840

Thank you for everyone again, and in this session I will use R and the ArcGIS

398

00:47:29,840 --> 00:47:37,920

ArcMap to show you how to process the data and how the data looks like. Again, I'm using R - I don't

399

00:47:37,920 --> 00:47:44,480

know how much you know R, but if you don't have any idea about how about R

400

00:47:44,480 --> 00:47:50,400

try to understand the concept. The code and the data, all data, will be available on

401

00:47:51,440 --> 00:47:58,960

on your request. So, if you want to follow my code please request it through the

402

00:47:58,960 --> 00:48:06,320

Urban Big Data Centre website. And I hope you can also have a recorded version of this seminar.

403

00:48:06,320 --> 00:48:13,600

About the code, there are several ways to build a R code. So, I'm not saying that my code

404

00:48:13,600 --> 00:48:20,240

is the most efficient one, you can use your own code for the same purpose so that's

405

00:48:20,240 --> 00:48:27,840

something that I want to talk before starting. So, I hope you can all see my folder and

406

00:48:27,840 --> 00:48:35,200

R Studio and this is the data you will get if you request the data from UBDC.

407

00:48:35,200 --> 00:48:43,600

Glasgow 2016, January 1st to December 31st ride edges. That is the cycling data from Strava data.

408

00:48:44,320 --> 00:48:50,160

And this is the Glasgow city boundary. Why we need this one, I want to show you

409

00:48:50,160 --> 00:48:57,600

later. And if you look at the folder you will see this one, this is a spatial data.

410

00:48:57,600 --> 00:49:06,080

This includes all the spatial information for all edges - the link, the row, segment - of the Strava data

411

00:49:06,080 --> 00:49:12,800

and this is the data you will get. Whole Strava data depends on different levels of aggregation

412

00:49:12,800 --> 00:49:21,200

and this one is the hourly aggregation per each link, each edges, for example. So that is what

413

00:49:21,200 --> 00:49:25,360

we are going to use, and this is the data format that you will get. Then let's look at the

414

00:49:25,360 --> 00:49:34,480

data in detail. First the spatial data. This is ArcMap. You need a license to use it but you can

415

00:49:34,480 --> 00:49:41,440

download the QGIS for free and that's very similar. And you can do the same thing that

416

00:49:41,440 --> 00:49:50,800

I'm doing here in QGIS. So, if you want to follow my instruction then you can download the QGIS.

417

00:49:51,680 --> 00:50:00,720

First, so this is the one that I show you the data folder and see Glasgow shapefile. you need all

418

00:50:00,720 --> 00:50:09,600

this one, this data set to get the one shapefile. If you click one you see this one.

419

00:50:10,480 --> 00:50:16,480

That is the data from Strava. They use some secure boundary, and they clipped all the edge

420

00:50:16,480 --> 00:50:23,760

information and extract those Strava data. But if you look at this one in the attributes,

421

00:50:24,800 --> 00:50:28,960

right click attributes, this is the information that inside this shapefile.

422

00:50:28,960 --> 00:50:36,800

The spatial data there is id, so if I select one link

423

00:50:39,040 --> 00:50:47,040

it will be highlighted and if you look at the date on that one it's id is 747718.

424

00:50:47,040 --> 00:50:54,320

That is the id that we need to use to merge Strava actual data with this edge data.

425

00:50:54,320 --> 00:51:01,200

Because Strava data is Excel file a CSV or DBA file and then this is a special location of

426

00:51:01,200 --> 00:51:09,680

the edges. So, we need to merge them later. There are many information about the edges, the rows, segments

427

00:51:09,680 --> 00:51:16,080

and here is kilometres, which is the length of the edges. I want to make sure, because

428

00:51:16,080 --> 00:51:22,640

it depends on the map, but the length of the edges are different so we may need this

429

00:51:22,640 --> 00:51:28,880

information to calculate total cycling distance and the location of the edges. So, that is one.

430

00:51:31,600 --> 00:51:36,880

The thing is, you may not want to use all data sets, you may want to only the edges

431

00:51:36,880 --> 00:51:45,760

in a city. Right, that's possible. So, I download another data set which is

432

00:51:46,720 --> 00:51:53,520

Glasgow City boundary from the Scotland website.

433

00:51:55,520 --> 00:52:04,000

And I think you can get a link from UBDC about that. We will provide it if you

434

00:52:04,000 --> 00:52:10,800

request it. And if you look at this one, this is a city boundary of Glasgow.

435

00:52:11,520 --> 00:52:18,320

So, I want to only extract and select the edges inside the city boundary for the first purposes.

436

00:52:18,960 --> 00:52:25,520

And how can I do that? Here I use the clip function in analysis.

437

00:52:26,640 --> 00:52:33,680

So, if I click this one, you see this one, and input features I press Strava edge data.

438

00:52:35,280 --> 00:52:43,680

Right, and then clip features I put the city boundary data and then here I just define the

439

00:52:43,680 --> 00:52:50,960

name of my final result Glasgow clip shapefile. You can define

440

00:52:50,960 --> 00:52:58,560

the location and then put the name here. So, if I click here, ok, then what you've got is

441

00:52:59,520 --> 00:53:06,560

this data. You see there are the different colours, so I will only show that

442

00:53:06,560 --> 00:53:14,560

one. This is the final result. I only selected the edges inside the city boundary, and it has the

443

00:53:14,560 --> 00:53:23,440

exact same data. Right, ID and kilometres and x y coordinates, but I only selected the edges inside

444

00:53:23,440 --> 00:53:33,200

the City of Glasgow boundary. Ok, and then save as a Glasgow clip. So, if I look at my folder again,

445

00:53:33,920 --> 00:53:41,840

Glasgow clip, there's all different shapefile information and in DBF file that includes the data.

446

00:53:42,880 --> 00:53:52,160

Like this data, it's a DBA file. The same here, right. So now we know what edges are inside the

447

00:53:52,160 --> 00:53:59,680

Glasgow area, right, and then we will use the folder to process the

448

00:53:59,680 --> 00:54:08,800

data. So, I will close this one. That's all I need for ArcMap. So, this is R Studio.

449

00:54:08,800 --> 00:54:15,040

Everything is free, you can download it and then if you want to use the specific

450

00:54:15,040 --> 00:54:22,880

function in R you need to download and install the packages and then import the packages. And this is

451

00:54:22,880 --> 00:54:30,000

how I import the packages library. I need tidyverse and lubridate and others to process data.

452

00:54:31,600 --> 00:54:42,080

So, I run this one here. So first I import all the library and first in R I need to tell R

453

00:54:42,080 --> 00:54:48,480

where my data are stored, right. So, this is the folder.

454

00:54:49,760 --> 00:54:58,000

So, in the user my ID and Documents and 2020 Webinars Strava and Glasgow edges. That's the folder

455

00:54:58,640 --> 00:55:09,520

where my data are stored. So here, right this one. So, see this is the path of my folder and this

456

00:55:09,520 --> 00:55:18,000

is the one. So, I define in R where my data exists. So, I run this one and now the R knows where

457

00:55:18,000 --> 00:55:28,960

my data are and I use read CSV to import the data. Again, this CSV file is hourly aggregation

458

00:55:28,960 --> 00:55:39,920

of the Strava activities. And then I saved it as Strava, right. I did it very quick. And then

459

00:55:39,920 --> 00:55:47,840

let's look at what information are in the Strava data. So, I use the summary function here.

460

00:55:49,680 --> 00:55:56,320

And see here, first there are almost 2 million observations. That's a lot, right?

461

00:55:56,320 --> 00:56:02,000

2 million observations here. And there are 14 variables. What are the 14 variables?

462

00:56:02,000 --> 00:56:09,600

This is the 14 variables inside the Strava. Edge ID - it's not ID it's Edge ID, so we need to,

463

00:56:09,600 --> 00:56:16,480

if we want to merge this spatial data with this Strava count data we use the ID from the

464

00:56:16,480 --> 00:56:22,880

shapefile and also Edge ID from the Strava because they are the same. It has year

465

00:56:23,600 --> 00:56:34,080

2016 and day 1 to 366 because there are 366 days in 2016. Hours 1 to 23, so one day

466

00:56:34,720 --> 00:56:42,400

and then minutes and athlete count is the number of Strava users on their roads at a particular time.

467

00:56:43,200 --> 00:56:47,920

And reverse means there are two directions on the roads, so they have a two direction

468

00:56:47,920 --> 00:56:55,760

one. And activity count is a cycling trip, number of cycling trips on that edge. And the reverse is

469

00:56:55,760 --> 00:57:02,000

again, there are two directions. And total is the sum of this one and this one for the activities.

470

00:57:02,000 --> 00:57:06,640

And then activities is the sum of activity count and reverse activity counts, that's the one.

471

00:57:07,680 --> 00:57:14,400

There's time information, also commute count. There in the app you can, after you finish

472

00:57:14,400 --> 00:57:19,840

your whole journey, you can indicate whether this is a commuting trip or not. So, if they indicate as

473

00:57:19,840 --> 00:57:24,160

commuting, this is the information about the commuting count and that is the

474

00:57:24,160 --> 00:57:32,720

information we can use to separate commuting and non-commuting trips. I hope this makes sense. And then

475

00:57:34,960 --> 00:57:42,640

I will first here, but as I said this Strava data includes all Strava activity data for

476

00:57:42,640 --> 00:57:49,440

all edges with what we first see, so it takes a little time.

477

00:57:50,560 --> 00:57:56,880

So, this one, right, Strava data includes all the edge information.

478

00:57:58,800 --> 00:58:06,400

This one. So, what I need to only select the edge information, the Strava activity

479

00:58:06,400 --> 00:58:13,760

information, from edges that are inside the Glasgow boundary. So how can I do that? Here

480

00:58:14,400 --> 00:58:26,160

I read the DBA file so Glasgow clip 1 DBF. So that actually gives us all the ID of the

481

00:58:26,160 --> 00:58:33,680

edges which are within the Glasgow boundary. Does it make sense? So those are the edges inside the

482

00:58:33,680 --> 00:58:41,520

Glasgow boundary and this is the function, this is the command.

483

00:58:41,520 --> 00:58:48,240

I use the Glasgow data and then only select the ID and kilometres, the length of the edges,

484

00:58:48,240 --> 00:58:54,560

because that's the only information that I need from the spatial data. And I use this data

485

00:58:55,840 --> 00:59:00,560

and the inner join, use inner join. What is the inner join? There are different types of join but

486

00:59:00,560 --> 00:59:09,120

inner join means you only keep the data set that matched. So here I joined the data,

487

00:59:09,680 --> 00:59:18,000

the Glasgow clip data, ID and the Strava data - this one. The excel file,

488

00:59:18,000 --> 00:59:27,280

the CSV file by using ID from the Glasgow file, which is a shapefile, and Edge ID from the Strava data.

489

00:59:28,400 --> 00:59:35,840

So, I did it and inner join means only keep the observation that are matched. So

490

00:59:35,840 --> 00:59:43,120

I did it and Glasgow Strava you see now we have 1.4 million observations because we removed all

491

00:59:43,120 --> 00:59:52,160

the counts beyond the edges beyond the Glasgow city area.

492

00:59:53,440 --> 00:59:56,000

Is it clear, yeah? It's a little bit

493

00:59:56,640 --> 01:00:01,520

hard to see your faces, I cannot see your faces so it's very hard to see whether you understand

494

01:00:01,520 --> 01:00:07,920

it or not but I will keep doing that. We can have a Q&A session later. So now we have all the

495

01:00:08,480 --> 01:00:16,080

the cycling activity data, Strava data for edges within the Glasgow city boundary. I want to see

496

01:00:16,080 --> 01:00:23,520

the travel patterns, cycling patterns, and I also want to check the quality of the data by producing

497

01:00:23,520 --> 01:00:30,320

a map. If we produce a map, we can easily see whether the numbers, the data, makes sense or not.

498

01:00:30,960 --> 01:00:38,160

So, to do that, first thing, what I did here is I calculated the total count per edges

499

01:00:38,160 --> 01:00:47,760

for a whole year. The total annual counts per edges, road segments. Here I use the Glasgow Strava data here

500

01:00:48,320 --> 01:00:57,280

we joined it, right, and then group by ID. The reason is I want to calculate total cycling distance

501

01:00:57,280 --> 01:01:06,160

per ID per edge. So that's the thing I need to first group them by ID and then use summarise

502

01:01:06,160 --> 01:01:15,040

function to calculate sum of, here I use the total activity count, the total count, number of cyclists,

503

01:01:15,760 --> 01:01:23,600

cycling trips, and multiply by the length of the edges. Why I want to calculate the cycling distance,

504

01:01:23,600 --> 01:01:29,360

total cycling distance, and that's all, again the reason is the length of the edges are

505

01:01:29,360 --> 01:01:35,520

different, so if we just use a total cycling count, they may be a little bit confusing.

506

01:01:36,320 --> 01:01:43,600

So, we calculate the sum of this total activity count multiplied by km,

507

01:01:43,600 --> 01:01:51,360

kilometres, and then save it as the total distance and we calculate this total distance by each

508

01:01:51,360 --> 01:02:01,200

ID, each edges. So, I did it and then save it as a total count. So, let's look at the total count,

509

01:02:02,640 --> 01:02:17,280

So, Edge ID 105, the total cycling distance is 192 kilometres or 106 it's 494. So, we calculate total

510

01:02:17,280 --> 01:02:25,600

annual cycling distance by each edge, right. That's what I will do, what I do here. Then

511

01:02:26,160 --> 01:02:33,920

I want to show this total distance in a map. Then I can see where are the popular roads

512

01:02:33,920 --> 01:02:40,960

where other kinds of cycling, where the cycling activities happen, right. So, to do that

513

01:02:40,960 --> 01:02:47,440

I want to import the shapefile. It can be possible; you can use R to import the shapefile and

514

01:02:47,440 --> 01:02:54,640

produce a map in a nice way. To do that I need these two libraries, ggmap and sf.

515

01:02:54,640 --> 01:03:06,160

So, I imported it. They are here. Here I import the Glasgow clip shapefile. What is that? That is,

516

01:03:07,600 --> 01:03:15,520

again, it takes time to see the ArcMap. This data. Glasgow clip shapefile, this one. I imported

517

01:03:15,520 --> 01:03:26,480

it in R by using st_read, right. And so, I imported that data and then use

518

01:03:28,640 --> 01:03:35,920

inner join with total count. Total count includes the ID and then total cycling distance.

519

01:03:35,920 --> 01:03:40,000

This includes all the edge information inside the Glasgow boundary.

520

01:03:40,560 --> 01:03:47,040

Right? So, I merge these two and then save as a edge. So, I did it.

521

01:03:49,440 --> 01:04:00,240

So now I have edge data. Here there is 11,000 observations, the edges. And to import the base map

522

01:04:00,800 --> 01:04:09,120

we need to use st_bbox and edge which is the data set that we saved. We need to use the same coordinate

523

01:04:09,120 --> 01:04:14,640

system with edges. Inside the shapefile there's information about all the coordinate systems

524

01:04:14,640 --> 01:04:21,840

and the locations, right. And we need to define the boundary and to define the boundary we

525

01:04:21,840 --> 01:04:31,040

need to use st_bbox. So, we do that and if we see what is inside, it gives the four values about

526

01:04:31,040 --> 01:04:38,400

the boundary. To use the get_map function we need to change the column name as a left, bottom, right,

527

01:04:38,400 --> 01:04:45,840

top, that isn't fixed. So, if you do that and then see what happened, xmin changed the

528

01:04:45,840 --> 01:04:53,360

left, ymin changed the bottom. So, we changed it. This is required if we want to use a get_map function.

529

01:04:53,360 --> 01:05:02,080

And now I want to get the base map and save it as a Glasgow map. So, this is what I’m doing.

530

01:05:04,640 --> 01:05:08,080

Now I have a base map. You cannot see, because I didn't

531

01:05:08,080 --> 01:05:11,760

command the print, right. So, let's do it here.

532

01:05:14,400 --> 01:05:22,320

ggmap is how you print the map, so I said print the Glasgow base map

533

01:05:22,320 --> 01:05:29,360

here and then data is edges, this is the data, in data we have all the information of

534

01:05:29,360 --> 01:05:36,320

edges and then total count because we merged them. Total distance, cycling distance. I want to

535

01:05:36,880 --> 01:05:43,680

show the graph, each edges, but the thickness can change, the size can change depending on the

536

01:05:43,680 --> 01:05:50,480

total travel distance, cycling distance, right. So, we see if it's thicker then it means there are

537

01:05:50,480 --> 01:05:56,880

high levels of cycling activities. For the colour we use the Sienna1. You can change it

538

01:05:56,880 --> 01:06:04,560

to the black, blue, red, whatever you like, right. Scale size, for the travel

539

01:06:04,560 --> 01:06:12,640

total distance we need to define the breaks for the values, so we use 10, 50, 100, 150. I

540

01:06:12,640 --> 01:06:20,160

will show you what it means, and range is 0.123. You can change it. This defines the thickness of the

541

01:06:20,160 --> 01:06:26,720

breaks, depends on the breaks. And then we put the label, the title is total annual counting

542

01:06:26,720 --> 01:06:31,840

distance per edges and size is kilometres. Let's see what happens if we run this one.

543

01:06:36,240 --> 01:06:45,120

You see this nice graph. So, I will enlarge the plots. So, this is the kilometres our dependent

544

01:06:45,120 --> 01:06:55,120

variable and this is the breaks we defined - 10, 50, 100, 150 - and that has a different

545

01:06:55,120 --> 01:07:03,840

size, right. That's what we have here. And we use the sienna, this is a colour called Sienna1 and

546

01:07:03,840 --> 01:07:10,400

that's the graph. We see the river and we see a lot of cycling activities

547

01:07:10,400 --> 01:07:17,920

alongside the river and here. As you may remember in the previous session, we have this area and this

548

01:07:17,920 --> 01:07:22,640

area has a really nice cycling infrastructure, good cycling infrastructure and it totally makes

549

01:07:22,640 --> 01:07:28,800

sense. It's not hilly, it's very flat. And see the city centre there are quite

550

01:07:28,800 --> 01:07:34,880

good levels of activities. And then for whole real main roads there are some activities there.

551

01:07:34,880 --> 01:07:41,520

So, this shows that, oh yeah, the data looks reasonable and that's what we expected and that's the

552

01:07:41,520 --> 01:07:53,760

easiest way to see the travel patterns in your city, right. So that's what we did here. So, if we

553

01:07:53,760 --> 01:08:00,720

want to change the key variable you can change it here. It depends on your research but then now

554

01:08:02,800 --> 01:08:08,720

the main analysis that I'm going to do here is I want to examine the relationship

555

01:08:08,720 --> 01:08:15,920

between weather conditions and cycling activities, right. The relationship between weather conditions

556

01:08:15,920 --> 01:08:22,240

and cycling activities. So first I need to process, I need to calculate total cycling distance

557

01:08:22,240 --> 01:08:29,120

but for different purposes because covering cycling distance activities could be different

558

01:08:29,120 --> 01:08:35,760

from the non-commuting cycling patterns, right. So, we do that. How we do that? We want

559

01:08:35,760 --> 01:08:41,200

to calculate the total cycling distance per day because if we want to examine the

560

01:08:41,200 --> 01:08:48,240

relationship between weather conditions and the cycling activities we need to make our analysis

561

01:08:48,240 --> 01:08:55,760

unit as hourly or daily. But we decided to use daily here because weather changes by

562

01:08:55,760 --> 01:09:06,080

daily. So here I create total cycling data set but they use the Glasgow Strava data, which is this one,

563

01:09:06,080 --> 01:09:14,160

right. All the edge informations are there and also the activity Strava data, raw data. We now group

564

01:09:14,160 --> 01:09:21,360

by day because we want to calculate total cycling distance per day. So that's what we are doing here.

565

01:09:21,920 --> 01:09:29,440

and we use the summarise function to calculate total activities, means sum of total activity count

566

01:09:29,440 --> 01:09:37,040

multiplied by kilometres. So that's total cycling distance. For the commuting activities we use

567

01:09:37,040 --> 01:09:44,560

commuting count, right. In the Strava data we see there was committing count if the users tick

568

01:09:44,560 --> 01:09:53,440

this trip as commuting. And then kilometre again, the length of edges. For the non-commuting

569

01:09:53,440 --> 01:10:00,560

activities then how can we calculate, we can use the total activity count minus commuting count.

570

01:10:00,560 --> 01:10:07,200

That is the full non-commuting count, right, multiplied by kilometres. So then saved edge is

571

01:10:07,200 --> 01:10:14,560

non-commuting activities. So, this total activity means total cycling distance per day,

572

01:10:16,080 --> 01:10:24,880

right. So, we do that and let's see what is inside the total cycling data.

573

01:10:26,320 --> 01:10:33,680

See there is days 1 to 366. It's omitted, it

574

01:10:33,680 --> 01:10:44,240

just shows the first six rows and total activities in day one there is only 886

575

01:10:44,240 --> 01:10:51,120

kilometres. And for commuting none, because it's a new year, right. No one will

576

01:10:51,680 --> 01:10:58,320

work there at the time. And non-commuting there are some people. So, we calculate total

577

01:10:58,320 --> 01:11:05,840

cycling distance for commuting and non-commuting by day. Here,

578

01:11:07,440 --> 01:11:13,840

later, I want to use a nice graph and plot the date in a proper way, so I use

579

01:11:13,840 --> 01:11:21,520

as date function. By using this variable, so total cycling data,

580

01:11:21,520 --> 01:11:27,920

that's inside, there is a day variable and we use minus one because that's how we define the

581

01:11:27,920 --> 01:11:36,560

use the as date to match with this format. So, if I run this one, this command, what happens is

582

01:11:37,760 --> 01:11:44,880

let's look at the data again - it has a date! It's a nice format - year, month, and day.

583

01:11:45,440 --> 01:11:53,360

So, day one is January 1st, day 2 is January 2nd. So that's a very easy way to change the format

584

01:11:53,360 --> 01:12:03,600

from day to date. And here, let's do the sum. So, I calculated

585

01:12:03,600 --> 01:12:11,280

all the cycling distance, total distance, and then I want to check whether the data looks okay.

586

01:12:11,280 --> 01:12:17,600

There are many ways, but this is one way. Just do the summary and then let's look at

587

01:12:17,600 --> 01:12:26,560

the total activities which means total cycling distance by daily and the mean is 2,500

588

01:12:26,560 --> 01:12:34,880

kilometres for whole area but max is almost 10 times. That is weird. That's too large.

589

01:12:35,600 --> 01:12:45,520

Then there could be some issue. So, what I'm doing here is I just want to see which date. So

590

01:12:45,520 --> 01:12:53,520

using the total cycling data, and if the total cycling data, the total activity is greater than 23,000

591

01:12:53,520 --> 01:13:03,920

just plot them. That's what it is, what this command means. So, we run it and we have day

592

01:13:04,800 --> 01:13:13,120

255 and that is September 11th. And I googled it, what happened in Glasgow

593

01:13:13,120 --> 01:13:18,560

on September 11th and there was an annual Glasgow to Edinburgh bike ride event

594

01:13:18,560 --> 01:13:25,040

on that day. So that is an exceptional date and for the analysis it's better to remove it. So here

595

01:13:26,880 --> 01:13:34,400

I use the total cycling data, the same data, and filter so only select

596

01:13:34,400 --> 01:13:40,000

if the total activity is less than 23,000. So, if I do that

597

01:13:42,560 --> 01:13:52,400

now to the summary again, the max is 10k. You know, it's kind of better, much better, right. You

598

01:13:52,400 --> 01:13:57,520

can do more, if you are using the Strava data for your cities you can do more investigation.

599

01:13:57,520 --> 01:14:04,400

But I think that's okay for this tutorial. So now I process the data, I check the data,

600

01:14:04,400 --> 01:14:10,960

whether there are errors or exceptional days. And then I want to see the trend of the data

601

01:14:10,960 --> 01:14:17,520

because each time series data is 1 to 366. So, I want to see the trend of the whole activities.

602

01:14:18,320 --> 01:14:24,160

Here, I calculate the moving average. This is a good measure for showing the trend of the data.

603

01:14:24,160 --> 01:14:29,440

If we use the raw data there will be a lot of spikes, so that is very hard to see.

604

01:14:29,440 --> 01:14:36,080

But if you use the moving average it's nicer. Moving average means you average the whole past

605

01:14:36,080 --> 01:14:43,040

seven days and use that average as your value. So, I actually find this code from

606

01:14:43,040 --> 01:14:47,760

online, I mean there are a lot of R codes you can just google it if you don't know how to do it.

607

01:14:48,640 --> 01:14:54,800

So, this is how we make a function to calculate moving average. I will briefly explain here the

608

01:14:54,800 --> 01:15:00,880

concept. So, this is the new variable that we are going to create, and this is the loop function.

609

01:15:00,880 --> 01:15:08,960

So, let's assume that the i is the seventh of January, right.

610

01:15:08,960 --> 01:15:16,640

Then the variable, this variable, the value will be the mean of activity which we'll define later as a

611

01:15:16,640 --> 01:15:24,400

travel or total travel cycling distance or commuting cycling distance. i minus n, let's see,

612

01:15:24,400 --> 01:15:34,400

i we said it's 7 and n equals, what, six. We already predefined it, so it becomes 1 and i becomes 7.

613

01:15:34,400 --> 01:15:41,040

So, what it means is, this function means, let's make a mean of activity your variable 1 to 7.

614

01:15:41,680 --> 01:15:48,400

So past seven days you use that variables, that values, and then calculate the mean

615

01:15:48,400 --> 01:15:57,680

and put the value here. The seventh value of the average variable. Does it make sense? Yeah that is

616

01:15:57,680 --> 01:16:04,320

what I'm doing here. So, I calculate, I create ma function, moving average function, and

617

01:16:04,320 --> 01:16:12,560

here what I'm doing is use this ma function that I create and then use total cycling distance

618

01:16:12,560 --> 01:16:18,880

total cycling distance for commuting purposes and non-commuting purposes. This is the variables.

619

01:16:19,440 --> 01:16:27,120

and use this function and calculate moving average and save as a ma total, ma commuting

620

01:16:27,120 --> 01:16:35,760

and ma non-commuting, right. And use mutate. Mutate is a function that when I want to make new variables.

621

01:16:36,960 --> 01:16:44,400

So, I use the total cycling, again the final dataset we processed here and then save as

622

01:16:44,400 --> 01:16:50,880

total cycling again. So, I don't want to make a different data set, I want to keep this original

623

01:16:50,880 --> 01:16:58,800

one and adding more. So, when I do that, see what happens.

624

01:16:58,800 --> 01:17:05,760

This command just shows me the first 10 observations and this one. They want total

625

01:17:05,760 --> 01:17:10,960

activities, commuting activities, non-commuting activities, that's what we have. And date, that's

626

01:17:10,960 --> 01:17:19,360

what we originally have, right. Now there is a new variable ma total ma commuting and I think

627

01:17:19,360 --> 01:17:25,280

because the size is so small it doesn't really produce the other one but let me do it again.

628

01:17:28,400 --> 01:17:35,280

Yeah, so the ma non-commuting, so this is the moving average. So, look here

629

01:17:36,480 --> 01:17:38,240

there is no change until

630

01:17:40,720 --> 01:17:48,320

five, fifth observation. Because if you look at this one if, for example, i equal 5 then this is 5 minus

631

01:17:48,320 --> 01:17:57,200

6 because 6 is predefined it's -1. I cannot really calculate this number so when it becomes greater

632

01:17:57,200 --> 01:18:03,920

than or equal to zero it calculates the moving average. So, this moving average is the average of

633

01:18:05,840 --> 01:18:14,080

this. Seven [counts] actually six values. If you average this one there will

634

01:18:14,080 --> 01:18:22,480

be this number. You can check it later, I already did it. So, we calculate the moving average

635

01:18:22,480 --> 01:18:28,800

and also, we have raw data, raw total cycling distance and total cycling distance for commuting

636

01:18:28,800 --> 01:18:35,520

and non-commuting. That's the data we processed. Here I want to show the trend. I

637

01:18:35,520 --> 01:18:42,800

want to make a graph that's showing the trend. So, I use the data that we've just processed. One problem

638

01:18:42,800 --> 01:18:51,360

is, I want to see the whole three information - total travel cycling distance, total commuting and

639

01:18:51,360 --> 01:18:56,560

non-commuting and also moving average for total commuting and non-commuting. But that's very hard

640

01:18:56,560 --> 01:19:03,680

because the current format has, for example, these three different columns. If you look

641

01:19:03,680 --> 01:19:12,160

at the graph plot, the y, there's only one variable, right, so that's very hard. This format, we call it

642

01:19:12,160 --> 01:19:20,800

a wide format. Through the nice graph you need to transform this wide

643

01:19:20,800 --> 01:19:27,200

form to the long form. So how can we do that? It's very confusing, right, just hold on I will explain.

644

01:19:27,200 --> 01:19:36,400

This is a function that we can change the format. So here, column 2:4 means I only use the column

645

01:19:37,120 --> 01:19:43,760

2 total activities, commuting activities and non-commuting activities. 2, 3, 4.

646

01:19:44,480 --> 01:19:52,160

And then use the names to activity type, create new variable activity type and put the value as a

647

01:19:52,160 --> 01:19:58,720

total distance. This is a new name of the variable, but I only use these three columns. So

648

01:19:58,720 --> 01:20:11,040

let's see what happens if I just run this code. Now see this one. Now it's the same but for one day,

649

01:20:12,160 --> 01:20:19,040

January 1st, there are three rows, and each row has activity time, a type

650

01:20:19,040 --> 01:20:26,320

as a total commuting and non-committing. And the total distance here is the value

651

01:20:26,320 --> 01:20:38,720

of the original value 886 here, 0 here and 886 again here. So, I change the wide format to the long

652

01:20:38,720 --> 01:20:45,840

format. So, we have one key variable - total distance - and we know activity type, different activity type.

653

01:20:47,600 --> 01:20:54,880

Does it make sense? And we use the ggplot to plot the trend. x is a date

654

01:20:55,760 --> 01:21:06,240

1 to 366 and y is the total distance. And then we said oh let's print the points, each value,

655

01:21:07,040 --> 01:21:12,720

but the shape changed by activity type and the colour changed by activity type because for

656

01:21:12,720 --> 01:21:19,120

the activity types we want to have a different shape and colours, right. And then this is the label

657

01:21:20,160 --> 01:21:27,440

and again, scale bar the x axis is a date and I want to have this format -

658

01:21:27,440 --> 01:21:37,680

year, month, and day. That's a nice format. And geom_line means let's connect all the points by line.

659

01:21:38,800 --> 01:21:45,360

So, I will show you the final results, that's better to understand, for your understanding. So

660

01:21:45,360 --> 01:21:51,680

I just use the raw data, not a moving average, and this is what it is. The total commuting activities, it has

661

01:21:51,680 --> 01:21:59,200

a blue colour and square shape because we define here, right, different colour and different shape.

662

01:21:59,200 --> 01:22:05,680

And it has this pattern. The non-commuting green and triangle and commuting activities

663

01:22:05,680 --> 01:22:12,240

red and circle. So, we have a different shape for different activity type and also the colours.

664

01:22:13,040 --> 01:22:18,000

See kind of seasonality impacts. There are low levels of cycling activities

665

01:22:18,000 --> 01:22:27,520

during the winter, here, but high activity levels during the spring, summer, and autumn.

666

01:22:27,520 --> 01:22:32,640

There are some decreases because I think this is because of the holidays - there are

667

01:22:32,640 --> 01:22:39,360

not many students here. Also, people are taking their holiday. Again, this is not really easy to

668

01:22:39,360 --> 01:22:46,640

see the trend, so I use the moving average. It's the same command but I use now column six

669

01:22:46,640 --> 01:22:54,160

to eight because that's the moving average that we calculate 1, 2, 3, 4, 5, 6 so six, seven

670

01:22:54,160 --> 01:23:00,800

and eight, right. And the same thing, it's exactly the same code. So, if I do that

671

01:23:03,840 --> 01:23:11,040

you see a much nicer trend, right. That is a reason why people use the moving average. You can

672

01:23:11,040 --> 01:23:16,160

ignore this part because that's not really moving average. So, this is moving average. You can see the

673

01:23:16,160 --> 01:23:23,600

seasonality impacts and also kind of variations because it's weekdays and weekends.

674

01:23:23,600 --> 01:23:28,880

Also, weather can be the factors why there are such a big variation.

675

01:23:31,200 --> 01:23:38,080

So now we have processed the Strava data. We know we have all the total cycling

676

01:23:38,080 --> 01:23:45,680

distance, cycling distance for commuting trips and non-commuting trips by day. So, we processed data

677

01:23:46,320 --> 01:23:52,480

Now we need weather data, right, so then we can build a model to see the relationship between

678

01:23:52,480 --> 01:24:03,920

weather conditions and the cycling activities. So, there are two data sets that you can use. One is you can

679

01:24:03,920 --> 01:24:10,000

get your data from your local weather stations, that could be more comprehensive.

680

01:24:10,000 --> 01:24:16,640

But if you don't have it you can use these two libraries to obtain the weather data for your city.

681

01:24:18,400 --> 01:24:25,360

So first I want to calculate the length of day because, again, the length of day is very important

682

01:24:25,360 --> 01:24:32,880

for cycling activities and in Glasgow the length of the day changes significantly compared

683

01:24:32,880 --> 01:24:44,000

to winter and summer. So here I make sunlight data. First, I define

684

01:24:44,000 --> 01:24:51,360

the data frame the date. We need to ask them what date we need the data and then the location.

685

01:24:51,920 --> 01:25:00,720

So, I use the total cycling date information that includes the 2016 January 1st to 2016 December

686

01:25:00,720 --> 01:25:08,000

31st and then put as a date. So that's a data frame as in the data frame there's a date

687

01:25:08,000 --> 01:25:14,480

variable, which is defined like this one and we put the latitude and longitude of our city area.

688

01:25:14,480 --> 01:25:19,600

How can you do that? If you just google it, your city, Google will show you the latitudes and

689

01:25:19,600 --> 01:25:25,840

longitudinal information, you choose information, so you can just type it. That is our data

690

01:25:26,480 --> 01:25:34,720

and we use getSunlightTimes function to get the sunrise and sunset time, right. That's what

691

01:25:34,720 --> 01:25:41,680

we do here. And we mutate length of day as the difference between the sunset and sunrise.

692

01:25:43,040 --> 01:25:50,320

Then we can estimate the length of the day and we rename day date as a date no time because

693

01:25:50,320 --> 01:25:56,960

sometimes date has its own function, so it can be confusing, but it's not really necessary. And then

694

01:25:56,960 --> 01:26:04,640

we have a day; we want to have a day like 1 to 366 because it's easy to

695

01:26:04,640 --> 01:26:09,840

use for merging other data set by using date no time. Date no time now is

696

01:26:10,400 --> 01:26:18,000

January 1st, something like that. So, if we do that and let's look at the data.

697

01:26:21,680 --> 01:26:29,200

and this is the date no time. It's 2016 January 1st and this format, we have latitude and

698

01:26:29,200 --> 01:26:35,520

longitude information, sunrise information, sunset information. We have length of days, seven

699

01:26:35,520 --> 01:26:42,960

hours it's been increased, and day 1 to 366, that's what we are doing here. So now we calculate

700

01:26:42,960 --> 01:26:54,640

the length of the day for 2016. Now we need to get more detailed weather data, like precipitation,

701

01:26:54,640 --> 01:27:02,720

like wind speed and temperature. How can we do that? We can use the getMeta function to get

702

01:27:02,720 --> 01:27:09,120

station information for your cities. So here I also put the latitude and longitude for Glasgow

703

01:27:09,680 --> 01:27:20,560

and then get the metadata inside the code. We see different stations, it takes time, so

704

01:27:20,560 --> 01:27:29,040

these are the stations, weather stations in Glasgow. And you can use the code to select the

705

01:27:29,040 --> 01:27:37,680

station. Here we select Prestwick, the weather data, weather station

706

01:27:37,680 --> 01:27:46,560

in Prestwick Airport. The reason is it has a list missing value so we use that code, so we use

707

01:27:46,560 --> 01:27:53,680

import NOAA function. We define the station and then we set each year to 2016

708

01:27:53,680 --> 01:28:03,840

and only select date, wind speed, air temperature, precipitation, right. And the mutate date no time

709

01:28:03,840 --> 01:28:13,840

again as a date, which means 1 to 366 and then hour of day. That's what we do here.

710

01:28:16,160 --> 01:28:20,400

So, we only select wind speed, air temperature and precipitation.

711

01:28:22,640 --> 01:28:30,960

And then I look at the weather data plus three here, you see this is the wind speed,

712

01:28:30,960 --> 01:28:38,800

air temperature, precipitation, date no time and hour of day 0 to 23, right. That's what we

713

01:28:38,800 --> 01:28:46,080

are doing here. So now we have all the weather data and rather than

714

01:28:46,080 --> 01:28:54,000

using the original data set, we want to have a mean temperature, max temperature, and some of the

715

01:28:54,000 --> 01:29:00,160

whole precipitation level and min wind speed and max wind speed because that's the more important

716

01:29:00,160 --> 01:29:09,840

determinants of the cycling activities and that's what we do here. So here if we look at the

717

01:29:11,440 --> 01:29:18,800

weather daily, we calculate the mean temperature, max temperature based on the original data set.

718

01:29:20,720 --> 01:29:24,400

I think that's straightforward.

719

01:29:25,200 --> 01:29:32,560

Then we have weather condition data, and we have length of day data. The next one is we need to

720

01:29:33,360 --> 01:29:40,400

merge these two data sets so we can have full weather condition data. So how can I do that?

721

01:29:40,400 --> 01:29:46,000

I have weather daily data, which is the final data set for weather, and sunlight data

722

01:29:46,000 --> 01:29:56,080

where I calculate the length of the day and both of the data sets has a day 1 to 366.

723

01:29:57,280 --> 01:30:05,840

So, we use that variable to merge these two data sets. And now I have a final weather data set.

724

01:30:08,320 --> 01:30:15,520

It says all the mean temperature, max temperature, precipitation, min wind speed, max wind speed,

725

01:30:15,520 --> 01:30:22,320

latitudes, longitudes, and length of day. So that is my whole weather data.

726

01:30:23,120 --> 01:30:33,200

Now what I need to do is merge this weather data with cycling data. In the total cycling data set

727

01:30:33,200 --> 01:30:40,480

we already calculated this total cycling distance for commuting and non-commuting and also overall

728

01:30:40,480 --> 01:30:50,160

total. We arrange it by day because we want to order the

729

01:30:50,160 --> 01:30:59,200

data set by day. And then use inner join, same thing, we merge weather data and this Strava data

730

01:30:59,840 --> 01:31:06,880

and then merge them. But before I do that, I want to make sure we have for the weather data

731

01:31:08,320 --> 01:31:20,560

we have 362 days because there are some missing values, right. So, we do that and

732

01:31:20,560 --> 01:31:26,560

then for cycling we have all the information. We have total activity, which represents

733

01:31:26,560 --> 01:31:33,600

the total cycling distance per day and the commuting cycling distance and non-commuting cycling distance,

734

01:31:34,480 --> 01:31:41,920

right. And all the weather condition data, that's what we have. So that is the final stage. We have

735

01:31:41,920 --> 01:31:48,240

all the data processed, now we want to see the key dependent variable, how they are distributed.

736

01:31:48,240 --> 01:31:54,560

So here I plot the key dependent variables. Total activities,

737

01:31:55,120 --> 01:31:59,760

again, total cycling distance, total cycling distance for commuting and non-commuting.

738

01:32:00,720 --> 01:32:07,040

They are skewed, that's general, right. So, if we want to use some kind of regression model it's

739

01:32:07,040 --> 01:32:13,840

better to make as a normal. So, we took here the square root, we take the square root transformation

740

01:32:14,480 --> 01:32:23,280

of the key dependent variable and then see how it looks. Yeah, this one and this one

741

01:32:23,280 --> 01:32:28,880

is much better. For the commuting still it's not really ideal, so you need

742

01:32:28,880 --> 01:32:34,480

more investigation if you want to conduct a proper analysis.

743

01:32:34,480 --> 01:32:41,600

But for this tutorial let's just go with it. So here, although this is the time series data

744

01:32:42,400 --> 01:32:48,880

so the observations could be correlated, we assume that they are independent. So let's

745

01:32:48,880 --> 01:32:55,040

just run the linear regression model, which means that we

746

01:32:55,040 --> 01:33:01,440

assume that all observations are independent. And this is the square root of the total activities,

747

01:33:01,440 --> 01:33:08,560

is our dependent variable and this is all our weather condition variables, right. We do that

748

01:33:09,760 --> 01:33:17,760

and print the result. We see four weather conditions have very

749

01:33:17,760 --> 01:33:24,080

significant relationship with level of total cycling activities, cycling distance.

750

01:33:24,640 --> 01:33:32,160

So, precipitation has a negative relationship. It means if the level of precipitation

751

01:33:32,160 --> 01:33:40,000

increases, more rain, the activity level will decrease, that's what it means. Max temperature

752

01:33:40,000 --> 01:33:45,840

increases the level of cycling activities. It makes sense because in Glasgow

753

01:33:45,840 --> 01:33:53,680

max temperature even in summer is not really that high. Max wind, if the wind speed is high

754

01:33:53,680 --> 01:34:00,160

there are fewer cycling activities. That totally makes sense. Length of day, if the length of

755

01:34:00,160 --> 01:34:06,240

day increases, the total cycling distance increases and that also makes sense. These are all consistent

756

01:34:06,240 --> 01:34:14,400

with previous studies, right. I want to check the model result and then see the residual

757

01:34:15,200 --> 01:34:20,800

path from the model to check the model assumptions because linear regression model,

758

01:34:20,800 --> 01:34:27,680

there are several assumptions. And it looks ok, actually, there are not many clear patterns and

759

01:34:27,680 --> 01:34:36,400

then normal ggplot looks ok. However, again, there could be an auto correlation issue and then to

760

01:34:36,400 --> 01:34:44,960

test auto correlation between observation we did a Durbin Watson Test. And then see there are

761

01:34:44,960 --> 01:34:51,040

several lags that have a p-value less than 0.05, which means there are auto correlation

762

01:34:51,040 --> 01:35:00,080

issues. In that case you need to use time series data. So here I used auto arima function

763

01:35:00,080 --> 01:35:06,320

to use the time series models and then to do that I need a library forecast.

764

01:35:09,520 --> 01:35:15,920

And this is the result. You don't have to worry about this other extra coefficient that's about

765

01:35:15,920 --> 01:35:23,600

the time series coefficient. But this one, if you can compare the estimate this one with

766

01:35:23,600 --> 01:35:29,200

the previous one. And what we found is very consistent. Although there are some differences

767

01:35:29,200 --> 01:35:35,840

in terms of magnitudes, the level of significance and also the signs are very consistent so we can

768

01:35:35,840 --> 01:35:41,520

conclude that yes there are significant relationships between weather conditions

769

01:35:41,520 --> 01:35:48,320

and cycling activities, total cycling activities. If you want to examine the commuting

770

01:35:48,320 --> 01:35:54,480

cycling distance and non-commuting cycling distance, you can just change this variable, right, and that's

771

01:35:54,480 --> 01:36:02,880

the same. I checked the assumption of the time series model, but I still see some of the problems.

772

01:36:02,880 --> 01:36:07,360

This is beyond this tutorial, so I don't want to talk about the models but

773

01:36:07,360 --> 01:36:14,160

as a researcher you may want to try other approaches to fix the auto correlation issues. So

774

01:36:15,040 --> 01:36:21,520

sorry for the long tutorial. I think it's a little bit hard to explain without your reactions but I

775

01:36:21,520 --> 01:36:28,400

hope you can get something from my tutorial. Again, this code will be available based

776

01:36:28,400 --> 01:36:35,600

on your request. So, if you need this code and data please do apply through the UBDC website. Thank you

777

01:36:35,600 --> 01:36:42,240

very much. [Muir Houston] And thank you very much for that Jin. We only have a couple of questions. One of them

778

01:36:42,240 --> 01:36:48,960

was about the code, so you've answered that one already. One more question, when you correlate

779

01:36:48,960 --> 01:36:57,120

Strava with cordon did you use the Strava edges or the Strava nodes in brackets intersections? [Jin Hong] So

780

01:36:57,120 --> 01:37:03,440

we use edge data because the location of the cordon count, that's not really across the

781

01:37:03,440 --> 01:37:09,600

node. It's more likely between, for example, middle of the edges or something like that.

782

01:37:09,600 --> 01:37:14,400

But I think that's the same because Strava they record all the people who pass that

783

01:37:14,960 --> 01:37:21,440

point. So, we use edge information and correspond to the location of the cordon count.

784

01:37:23,760 --> 01:37:28,960

[Muir Houston] That's great. And just one more, which road networks are you working with, which is just

785

01:37:28,960 --> 01:37:34,800

the city of Glasgow I think is it Jin? [Jin Hong] Yes, that's the location that we use

786

01:37:34,800 --> 01:37:41,520

here. The whole data is a Glasgow one, so here, this area.

787

01:37:45,760 --> 01:37:49,600

The data you will get for this tutorial is the Glasgow area.

788

01:37:51,680 --> 01:37:57,840

[Muir Houston] Ok and just to remind everybody, as Jin has said, and I've posted the links to the

789

01:37:57,840 --> 01:38:05,120

data catalogue on the UBDC which gives information about gaining access to the data.

790

01:38:06,240 --> 01:38:14,320

And I've also put the link there for the free GIS software QGIS and the link to R which is

791

01:38:14,320 --> 01:38:24,320

also free and open-source code. So, I'd just like to thank everyone for attending and thank Jin for

792

01:38:24,320 --> 01:38:32,640

his presentation and workshop. And the recording of this will be on the UBDC

793

01:38:32,640 --> 01:38:37,760

website, but we will need to make sure it's ok for accessibility given new regulations

794

01:38:37,760 --> 01:38:44,080

about accessibility of online content. So once again, thanks very much for coming and keep an

795

01:38:44,080 --> 01:38:48,480

eye on the UBDC. We've got another three of these data dives over the next month.

796

01:38:49,040 --> 01:38:54,960

So please, if you're interested, sign up and register for these and hopefully we'll see some

797

01:38:54,960 --> 01:39:02,760

of you at these other sessions. So once again, thanks everyone. [Jin Hong] Thank you very much, bye. [Muir Houston] Bye.