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.
I see we have participants from Australia, Austria, Belgium, China, Germany, Iraq, the Philippines, Russia, Turkey, Ukraine, and the UK. Sorry if I've missed any of you.

As you will have seen this session is recorded and will be uploaded on the web in an accessible format at some point after the session. Details will be provided on the UBDC website.

Also, check out the website for other resources, including how to access other data and other training and events delivered by the UBDC. As we've mentioned, cameras will be turned off and microphones muted to aid privacy and also for bandwidth reasons. Please use the Q&A facility to ask questions. These will be collated, and responses will be provided in the Q&A.
session. In terms of the session structure, Jin will give a presentation for around 30 minutes,

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

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

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

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

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

delay. In this session, as Muir said, I will talk about crowdsourced cycling data,
that is Strava data, and as a researcher how we have used the data for cycling studies.

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

Here's a brief background about the studies. So, the large benefits of cycling have been well documented. It could reduce the auto dependency, therefore reduce the level of congestion and emissions. It could also improve the public health because people are doing physical exercise while cycling. In addition, if you look at the travel surveys from different countries, you will notice that a substantial amount of the automobile trips are short trips. That means their travel distance is between two and five kilometres. What does
that mean? This is quite a long distance for walking. However, this is a really reasonable travel distance for cycling. So, it implies that cycling can be a good alternative of the automobile.

So, because of this huge benefit and the potential, many countries have used their substantial resources to improve the cycling environment and also increase cycling. And this is the same for the UK. The active travel - walking and cycling - is one of the priorities for the National Transport Strategies in the UK. In Scotland we have a really ambitious vision. Transport Scotland wants 10 percent of journeys to be made by bicycle by 2020. This is really ambitious.
because now we have one to three percent at most. And the cities are responsible for achieving this.

So, again we are trying to promote cycling. In Glasgow, the local government have also introduced several measures and interventions to promote cycling. For example, they want to use the 2014 Commonwealth Games -

to increase the cycling. So, they provided several cycling infrastructure lanes before, during and after the Commonwealth Games. In addition, they also provide bike share programme Nextbike as you can see in the pictures. They are quite popular at this moment. However, as a planner if you
want to make better cycling plans, you really need to understand the cycling patterns, cycling behaviour,

and also, you need to understand how to use the proper data to evaluate the effectiveness of interventions. Unfortunately, these are very difficult because we don't have proper data. We have travel surveys. So, many people actually use travel surveys to examine travel behaviour.

However, there are only a small number of cyclists in most cities and although its representative sample does travel surveys, they only include a small portion of the people from the population. What happens is, at the end, you may end up with 40 or 50 people who cycled.
in your travel survey for a metropolitan area. Then that’s too small. You cannot really use that
data to build some model or to analyse detailed cycling activities. We also have a manual and
automatic count. So, for example, in Glasgow every year for two days they manually count how many
cycles in and from the city centre and in some cities they installed the automatic counters
to continuously measure the cycle activities. Again, these are very expensive hence infrequent. There
are only a small number of automatic counters in the city because again it's very expensive.
So, it's a really good ground to this data they are, but there are significant
limitations if we want to use this data to examine the cycling patterns, cycling behaviour.
Due to the technology improvements, now we have new forms of data and these data provide detailed cycling activities at the fine spatial and temporal scale. The Strava cycling app is one of them and I think it's one of the most popular cycling apps in the world.

And they use GPS to track cyclist's journeys, so they know exactly what time and where the Strava users are using cycling. So that is a really amazing data set and as time passes more people are using this app because it became popular. So, then what does that mean? The quality of the data will improve because there are more data.
In addition, the data are already being collected all over the world because everyone can
download the app. So, we can compare the same policy in different countries using
the same data format and we also can use the same methodological approaches
to different cases because the data structures are exactly the same. However, as a researcher we
really need to understand then what could be the potential weaknesses of these emerging
forms of data, which could influence our study. So, what are the weaknesses? The first thing
is representativeness. I guess you already know this one. For example, in the Strava case,
it's more likely young male people are more likely to use the Strava apps and they are more
experienced cyclists than the normal cyclists. So, if their cycling patterns are different from those of the normal cyclists then the research, the analysis, could be biased. So, we could get incorrect conclusions. So, there are many people recently who have tried to employ the advanced analytics methodology methods to correct the bias. There are also special variations.

As I said, most cities have only a small number of cyclists and among them only a small number of people actually uses Strava apps. So, if we look at the popular roads you may find some people who use the Strava apps, but if you look at the less popular roads although there are actual cyclists you may not have any Strava users. So,
it depends on where the place is. Some people argue that Strava data is more useful for urban areas where the level of cycling activities is high compared to the rural areas.

Lack of social demographic information, mainly because of privacy issues. We all know that the social demographic factors are very important determinants of the travel behaviour like age, gender, income, education level and so on. However, we don’t have that information.

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

This is actually very important, and I will talk about the Strava case later. Because the data
is owned by the private company, if they somehow change their regulation or their products
because

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

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

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

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

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

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

collecting the data, because several people are using the Strava apps it's cost saving they
don't really pay anything and it's very detailed cycling activity at the fine spatial and temporal scales and again a growing number of users. So, that could improve the quality of the data.

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

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

And I hope this will give you some kind of idea about how we use the data. And these are the three research questions that we aim to answer. The first one is can crowdsourced cycling data be utilised for cycling behaviour studies? This is about the quality of the data, whether the Strava data are good enough for studying the cycling activities. The second one is, if yes then
where commuting cyclists travel and what are the influential factors for their route choice?

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

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

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

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

By the way, these data are available in UBDC. You can get the data based on your request.
area level. Output area is a UK census area. It's pretty small. So, what it means is we know for each trip, where the trip starts - output area - and what output area the person travelled and where is the destination - output area. This is pretty nice data. Second one is a more detailed one - minute by minute link count. Link count is at the road segment. We know how many people cycled on a particular time, minute by minute, this is really detailed data. We also have information about waiting times at junctions and then aggregated demographic information - for example, age and gender for your city. They just give you an aggregated summary.
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 --> 00:15:04,320
As I said, there are a small number of cyclists and among them a small number of people are using

123
00:15:04,320 --> 00:15:10,720
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.
And because they are binning this data, they just lose huge information.

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

Now, they only provide hourly aggregation in the lowest level of aggregation, not a minute by minute. I think it’s because of that issue. So, that is a big issue. Again, if that happens again, the whole methodologies we have used before may not be available because as some methods use the more detailed temporal scale.

So, we used four years of the Strava data but at the time the unbinned data was available so we used unbinned data for our own study.
We also used a manual count of cyclists from cordon count carried out in Glasgow in the same time period. So, there are 38 locations and two days in general in September or per year. I will show you the location later. And we also use Glasgow cycle infrastructure data as you can see on the map that is the current infrastructure map. And this is a river, and we see there are really good nice cycling infrastructure alongside the river and here and other parts. So, we can see some areas like this one - east side of Glasgow - where it’s the most deprived area, there are not many good cycling infrastructure. So, we can see some kind of special inequality issues here.
So, for the research question one, about the quality of the data, how can we check the quality? We can check the ground truth data, which is a manual count data, with Strava data.

So, in the map you can see the 38 locations and that's the locations where people check the number of cyclists. And we know the location and time of the number of actual cyclists and we know exactly time and location from the Strava data. So, we compare them.

That's how we do that. So, there are two types of analysis. One is a correlation analysis and the other one is simple linear regression model. The equation shows the simple linear regression model.

\[ y \] represents the number of cyclists from the cordon count, so ground truth data.
And x Strava means the number of Strava cycling trips with a simple linear regression model. And

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

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

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

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

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

we use the 2016 Strava data but OD matrix. We can construct OD matrix
because we know the output area of the origin and destination for each trip. So, we compare

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

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

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

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

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

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

Strava data and we calculate the total number of Strava trips per output level per month.
It's monthly total. Why? As I will show you later, as we increase the level of aggregation the data quality becomes much better. So, monthly or average we think is really good enough for this study. We used output area as analytical units. Then we use the fixed effect poisson panel regression model as you're shown in the equation. So, the dependent variable is number of cycling trips in area i, output area i in month t. This is the four years of data, so panel data. And we have four infrastructure, you can see on the map. And then this one is interesting because the longest one connects the suburban area to the city centre. Here is the city centre.

So, this is the result for research question one - a correlation analysis and simple linear
regression model. We have different levels of aggregation because we think if we aggregate more and more, the data become larger and larger, we will have more counts so we may have a better kind of result. So, the lowest level of aggregation is hourly and the last one is the two days because we only have two days of cordon count data. And correlation, even hourly, we find point seven eight almost 0.8. This is very high. It gives us a positive signal. For the two days of aggregation we have almost 0.9, which is a really good kind of indicator. When we look at the linear regression model result, the estimates
for the Strava trips our independent variable is positive and very significant

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

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.

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

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

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

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

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

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

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

bottom because again there are only small number of people who cycle. But if we aggregate, we see
more clear pattern of the linear regression and noise becomes smaller. So, based on this result we concluded that, yes, it could be good proper data for cycling studies.

For the second research question, we compare the short routes and actual routes. The graph on your left panel it's actual Strava data, so that's where people really cycled. So, we see alongside the river there is some concentration, they are popular. That makes sense because amenity is one of the main factors for cycling and also safe cycling lane infrastructure. For the shortest paths you see more evenly distributed kind of patterns. And that also makes sense but
it's already hard to see. So, to make a better vision we calculate the difference between them.

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

cycling trips happened on these roads compared to predicted ones. And the black means they are unpopular, less popular, and the thickness means the size of difference. So, we see alongside the river where there are good cycling infrastructure, there are really many people cycling.

But around the city centre area there are some red lines and black lines, but they are very thin. It means their difference doesn't really get different. That totally makes sense.

Right? So, then we see based on our local knowledge, we want to figure out why some roads are less
popular. And here's one example this is one from the southside of the city centre and other

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

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

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

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

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

studies, in particular studies, as a barrier of the cycling. So, we see yeah, maybe that's the reason.
Another one, this is the east part of Glasgow

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

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

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

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

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

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

For the research question three, we use the model and these are the four investments. And we
remove the time trend and then see how the total monthly count changed and the line here, you see the different location of the line, that's when the new infrastructure was open to the public. Right? And we see the three of them have a pretty positive increase after the new infrastructure were opened. However, this one the Routes to Cathkin 1, which actually connect the city centre to the suburban area, the longest one, has kind of a negative trend. So, after these basic stats we jump into the model and then these are the research from our model. We first measure the overall effect, and you can see it's there is a positive impact and it's about eight percent increase. You
take the exponential to interpret the result. But the P-value is 0.08 which is greater than 0.05 so it's not really statistically significant. However, if you examine the separate effects, we notice that three among four have very positive impacts and they are very significant.

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

and the right one, the first one, Routes to Cathkin 1 has a negative kind of impact compared to the other output area where there's no infrastructure. So, that is the reason why we didn't really get the significant result for the overall effect. And these three new infrastructure, they are close to the city centre area and they include the segregated
lanes. So, they could provide some kind of policy implication. If you want to get a short-term impact
then better to build more cycling lanes maybe near the city centre where the
level of cycling activities is high and also the segregated lens could be important. So, this is
the end of the slides. If you want to look more in detail about each of the studies and
methodologies
you can look at these three papers. They are open access, so you can download for free
and this slide will be available for you, so you can get the full information. Thank you very much.

[Muir Houston] Thank you very much Jin. We have some questions here. If I read them out one by one Jin, you
could maybe try and answer them? So, the first question is, and I think you touched on this.

How much Strava use the apps for commuting rather than the kind of lycra warriors that we see racing about the city that you mentioned, how many are kind of normal commuters if you like?

So, in our case it was almost five percent of people are among the whole cyclists, five percent of the people are using the Strava apps and that’s for our city but it depends on city by city. [Muir Houston] Ok, thanks for that. Now probably there might be some answer in the link you gave for your papers, is there in terms of cycling research case studies a journal?
Is there transport planning or any other kind of common journals that you tend to publish in Jin? [Jin Hong] Yeah, so we published several papers in several leading transport journals. Transport Geography, Transportation and then Environment Planning A and B. So, we also published some paper in Transport Research part A, which is a top transport journal. So, you can find more information on my website or on the UBDC website. [Muir Houston] That's great thanks. And now the next one: have you tried to identify trips in cycling mode from other mobile phone data? For example, Telefonica or Voda UK instead of Strava. [Jin Hong] No, actually we haven’t because

now we are trying to get some mobile phone data, Urban Big Data Centre, we try to buy
some mobile phone data, then we couldn’t do it. That is actually the next step that we want to do because we want to also look at the other data sources. So, currently we are planning to buy some mobile phone data and we will also plan to conduct a survey with the apps so we can check the people's location and their trips. So, then we can use this data to detect mode choice from the mobile phone data. So, we hope we can have another session later using these new forms of data. [Muir Houston] Good stuff. Now another question here specifically about somebody who maybe knows Glasgow. In terms of popularity of routes for cycling, was consideration given
to the physical and environmental factors e.g. topography or gradient? I think what they're

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

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?

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

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

00:31:56,000 --> 00:32:02,240
alongside the river which is very nice infrastructure it provides and very flat.

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.

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

May I ask how much Strava charged for the data and what is the attitude of people in Strava towards research projects or are they very commercial-focused? [Jin Hong] Yeah, this is the one that we may ask to Andrew who’s the administration manager in UBDC, senior manager. I think we somehow spent 60k to buy the data for the Glasgow area for several years. But that depends on the license because, as you see, we purchase the data and so everyone who wants Strava data they can request and they can get it for free. So, if we are interested in the Strava data for
Scotland especially, you can get the whole years of data from Urban Big Data Centre. And

they are really eager to engage with academics. I want to say because we also have several conversations with them and they all know our studies. Sometimes they blog our studies.

The problem is the current situation about safety. So, they try to change their data products and this is because its associated with running their companies.

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

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

[Jin Hong] There are many actually. There are books and there are papers, but you need to really learn
the basic statistics model inside the traffic assignment model and also you need to run, if you want to only know the traffic assignment model you can just use, there are already code available online and also if you the easiest way is to use the travel demand models software TransCAD or VISUM. Then you can easily actually do that if you have the data. So yeah there are a lot of books if you can just google it or you can find several books or papers.

[Muir Houston] And a question about socio-economic variables - do you use that much Jin? [Jin Hong] No,
again, this is not actually included in the Strava data. They only provide, for example in the city of

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

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

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

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

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

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

circumstances it's easier to encourage people to cycle or walk more? [Jin Hong] Yeah, I think so and actually
interestingly we have one draft paper that examines the cycling patterns after COVID-19 lockdown in the UK and you

we saw a significant increase in terms of cycling activities. And yes, I think that’s because of the current situation as well as the new situation, like for bike share programmes.

They provide kind of free rides and also there are a lot of people who bought a bicycle. So yes, I think so. [Muir Houston] And the next one Jin, what proportion of all cyclists use Strava? Or to put it another way, how does your sample size for Strava data compare to the sample sizes for manual count data?
[Jin Hong] That's so great. So again, as I said, I think that's the same for the first question.

In our case it is about five percent, but again it varies city by city. So that's the reason why I said in one Strava user you could represent 25 people, actual cyclists, or sometimes, it depends on the city, it could be like 200 actual cyclists. And also, it depends on the location. It's urban area versus the rural area because, again, the spatial variation.

[Muir Houston] And another one. Could average speed be used to differentiate between the sports cyclists and the commuters maybe? [Jin Hong] I think the easiest one is using the time. If commuters, there are certain times that they are using the cycles like am and pm peak. That's better I think in terms
of separating the commuter trips and non-commuter trips and in Strava data they actually indicate whether they are commuting or not. So, if we are using the Strava data it's easy. If we are using other data sources, it's better to use the time, am peak and pm peak, to separate the commuting trips.

[Muir Houston] Any methods for correcting the self-selection bias inherent in Strava data? [Jin Hong] So, there are currently several papers. They use some statistical models and use other built environments or other data sets and build some models to correct this bias. And there are also some studies, they are conducting their own survey and ask people whether they are
using Strava apps or not and then their activities, patterns. And also, they

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

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

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

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

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

understanding the why as well as the what. Have you worked with anyone in adding that
type of data through, for example, a bespoke app? [Jin Hong] Currently for the cycling, no. It was very hard to

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.

00:39:29,360 --> 00:39:36,160
That was the main one, but again we are trying to purchase some mobile phone data

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

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?

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.

00:39:57,360 --> 00:40:04,000
That's traditional survey data and not really the new forms of data like apps.

00:40:05,840 --> 00:40:11,120
But we do have another data set that we have at Urban Big Data Centre is
iMCD survey which actually Muir just mentioned. There are survey, so it's a representative sample of the whole metropolitan area of Glasgow. And also, some 300 people carried a GPS life logging device for one week. So that's another data set we have. We have their travel diary and we have their GPS trajectory and also life logging picture data.

So, if we may use those kinds of data together for the research to get more information. [Muir Houston] And actually Professor Lido, who works on that data, is giving one of the data dives, I think it's next week sometime. It'll be on the website, the details of that,
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 accident information? [Jin Hong] We haven't done it, but we know there are studies and yes that's totally possible. We can also look at in Scotland and the UK SIMD, we know these crime rates of the areas so we can also see how they range between the crime rates and the cycling activities. And also look at the location, as the person asked. [Muir Houston] And just in terms of your work with Glasgow City Council Jin, are they quite receptive to use this data to help guide their cycling policies and plan infrastructure, for example new cycling lanes and stuff like that? [Jin Hong] They are very supportive all the
infrastructure data they provide us for our own research and we also provide our research to them.

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

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

are happy to collaborate with us.

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

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

[Jin Hong] Yes, so that is a reason why people try to correct the bias actually, by
using other data sources. However, as we showed in the study, the general pattern is

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

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

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

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

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

for each trip the cycling routes, but that doesn't guarantee the same person has
the same id, they don't have that kind of thing. Because, yes, if we can do that, that's a very serious

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,

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.

00:43:57,280 --> 00:44:04,480
[Muir Houston] And a question here about, I know UBDC is doing some work on

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

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

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

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

question for this session I think, could you use Strava to perhaps plan the integration of the bicycle as transport in cities where the bicycle use is just emerging? So maybe a link back to this kind of covid restrictions and a lot more increase in people cycling in cities that don’t have the infrastructure. Could Strava or something like that be helped to try...

[Jin Hong] Yes, I think so because the first thing that you need to do is you need to understand the cycling patterns because there is reason why people use certain roads. But if you can know the whole picture at once, in a simple way, that will really help you to make a better plan.
So yes, so that’s another benefit of the Strava apps because that’s available for everywhere in the world. So, you can, for example, there are some studies they’re trying to find out, examine the cycling patterns in Africa where there’s no real data and no infrastructure.

But I think that’s really one potential benefit of this kind of data. You can easily see the cycling patterns although there could be some bias, but you can easily see what roads are popular and why. [Muir Houston] That’s great Jin, thank you very much. That’s all of the questions for this session. Now I think we’re going to have a break before the more
practical session, so on my clock it's 10:55. If we say 11:15 we'll reconvene for the second

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

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.

00:46:48,560 --> 00:47:00,960

[Muir Houston] Hi folks, welcome back to everyone. I hope you've all joined us again as we start this second session, which is a more hands-on practical and

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

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

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

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.
Thank you for everyone again, and in this session I will use R and the ArcGIS ArcMap to show you how to process the data and how the data looks like. Again, I'm using R - I don't know how much you know R, but if you don't have any idea about how about R try to understand the concept. The code and the data, all data, will be available on your request. So, if you want to follow my code please request it through the Urban Big Data Centre website. And I hope you can also have a recorded version of this seminar. About the code, there are several ways to build a R code. So, I'm not saying that my code
is the most efficient one, you can use your own code for the same purpose so that's

something that I want to talk before starting. So, I hope you can all see my folder and

R Studio and this is the data you will get if you request the data from UBDC.

Glasgow 2016, January 1st to December 31st ride edges. That is the cycling data from Strava data.

And this is the Glasgow city boundary. Why we need this one, I want to show you

later. And if you look at the folder you will see this one, this is a spatial data.

This includes all the spatial information for all edges - the link, the row, segment - of the Strava data

and this is the data you will get. Whole Strava data depends on different levels of aggregation
and this one is the hourly aggregation per each link, each edges, for example. So that is what we are going to use, and this is the data format that you will get. Then let's look at the data in detail. First the spatial data. This is ArcMap. You need a license to use it but you can download the QGIS for free and that's very similar. And you can do the same thing that I'm doing here in QGIS. So, if you want to follow my instruction then you can download the QGIS. First, so this is the one that I show you the data folder and see Glasgow shapefile. you need all this one, this data set to get the one shapefile. If you click one you see this one. That is the data from Strava. They use some secure boundary, and they clipped all the edge
information and extract those Strava data. But if you look at this one in the attributes,

right click attributes, this is the information that inside this shapefile.

The spatial data there is id, so if I select one link

it will be highlighted and if you look at the date on that one it's id is 747718.

That is the id that we need to use to merge Strava actual data with this edge data.

Because Strava data is Excel file a CSV or DBA file and then this is a special location of

the edges. So, we need to merge them later. There are many information about the edges, the rows, segments
and here is kilometres, which is the length of the edges. I want to make sure, because

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

00:51:22,640 --> 00:51:28,880
information to calculate total cycling distance and the location of the edges. So, that is one.

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

00:51:36,880 --> 00:51:45,760
in a city. Right, that's possible. So, I download another data set which is

00:51:46,720 --> 00:51:53,520
Glasgow City boundary from the Scotland website.

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

00:52:04,000 --> 00:52:10,800
request it. And if you look at this one, this is a city boundary of Glasgow.
So, I want to only extract and select the edges inside the city boundary for the first purposes.

And how can I do that? Here I use the clip function in analysis.

So, if I click this one, you see this one, and input features I press Strava edge data.

Right, and then clip features I put the city boundary data and then here I just define the name of my final result Glasgow clip shapefile. You can define the location and then put the name here. So, if I click here, ok, then what you've got is this data. You see there are the different colours, so I will only show that one. This is the final result. I only selected the edges inside the city boundary, and it has the
exact same data. Right, ID and kilometres and x y coordinates, but I only selected the edges inside the City of Glasgow boundary. Ok, and then save as a Glasgow clip. So, if I look at my folder again, Glasgow clip, there's all different shapefile information and in DBF file that includes the data. Like this data, it's a DBA file. The same here, right. So now we know what edges are inside the Glasgow area, right, and then we will use the folder to process the data. So, I will close this one. That's all I need for ArcMap. So, this is R Studio. Everything is free, you can download it and then if you want to use the specific
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
of the Strava activities. And then I saved it as Strava, right. I did it very quick. And then

let's look at what information are in the Strava data. So, I use the summary function here.

And see here, first there are almost 2 million observations. That's a lot, right?

2 million observations here. And there are 14 variables. What are the 14 variables?

This is the 14 variables inside the Strava. Edge ID - it's not ID it's Edge ID, so we need to,

if we want to merge this spatial data with this Strava count data we use the ID from the

shapefile and also Edge ID from the Strava because they are the same. It has year

2016 and day 1 to 366 because there are 366 days in 2016. Hours 1 to 23, so one day
and then minutes and athlete count is the number of Strava users on their roads at a particular time.

And reverse means there are two directions on the roads, so they have a two direction one. And activity count is a cycling trip, number of cycling trips on that edge. And the reverse is again, there are two directions. And total is the sum of this one and this one for the activities.

And then activities is the sum of activity count and reverse activity counts, that's the one.

There's time information, also commute count. There in the app you can, after you finish your whole journey, you can indicate whether this is a commuting trip or not. So, if they indicate as
commuting, this is the information about the commuting count and that is the

information we can use to separate commuting and non-commuting trips. I hope this makes sense. And then

I will first here, but as I said this Strava data includes all Strava activity data for

all edges with what we first see, so it takes a little time.

So, this one, right, Strava data includes all the edge information.

This one. So, what I need to only select the edge information, the Strava activity

information, from edges that are inside the Glasgow boundary. So how can I do that? Here

I read the DBA file so Glasgow clip 1 DBF. So that actually gives us all the ID of the
edges which are within the Glasgow boundary. Does it make sense? So those are the edges inside the

Glasgow boundary and this is the function, this is the command.

I use the Glasgow data and then only select the ID and kilometres, the length of the edges,

because that's the only information that I need from the spatial data. And I use this data

and the inner join, use inner join. What is the inner join? There are different types of join but

inner join means you only keep the data set that matched. So here I joined the data,

the Glasgow clip data, ID and the Strava data - this one. The excel file,
the CSV file by using ID from the Glasgow file, which is a shapefile, and Edge ID from the Strava data.

So, I did it and inner join means only keep the observation that are matched. So

I did it and Glasgow Strava you see now we have 1.4 million observations because we removed all

the counts beyond the edges beyond the Glasgow city area.

Is it clear, yeah? It's a little bit hard to see your faces, I cannot see your faces so it's very hard to see whether you understand

hard to see your faces, I cannot see your faces so it's very hard to see whether you understand

it or not but I will keep doing that. We can have a Q&A session later. So now we have all the

the cycling activity data, Strava data for edges within the Glasgow city boundary. I want to see
the travel patterns, cycling patterns, and I also want to check the quality of the data by producing a map. If we produce a map, we can easily see whether the numbers, the data, makes sense or not.

So, to do that, first thing, what I did here is I calculated the total count per edges for a whole year. The total annual counts per edges, road segments. Here I use the Glasgow Strava data here we joined it, right, and then group by ID. The reason is I want to calculate total cycling distance per ID per edge. So that's the thing I need to first group them by ID and then use summarise function to calculate sum of, here I use the total activity count, the total count, number of cyclists,
cycling trips, and multiply by the length of the edges. Why I want to calculate the cycling distance,

total cycling distance, and that’s all, again the reason is the length of the edges are
different, so if we just use a total cycling count, they may be a little bit confusing.

So, we calculate the sum of this total activity count multiplied by km,
kilometres, and then save it as the total distance and we calculate this total distance by each

ID, each edges. So, I did it and then save it as a total count. So, let’s look at the total count,

So, Edge ID 105, the total cycling distance is 192 kilometres or 106 it's 494. So, we calculate total

annual cycling distance by each edge, right. That’s what I will do, what I do here. Then
I want to show this total distance in a map. Then I can see where are the popular roads

where other kinds of cycling, where the cycling activities happen, right. So, to do that

I want to import the shapefile. It can be possible; you can use R to import the shapefile and produce a map in a nice way. To do that I need these two libraries, ggmap and sf.

So, I imported it. They are here. Here I import the Glasgow clip shapefile. What is that? That is, again, it takes time to see the ArcMap. This data. Glasgow clip shapefile, this one. I imported it in R by using st_read, right. And so, I imported that data and then use inner join with total count. Total count includes the ID and then total cycling distance.
This includes all the edge information inside the Glasgow boundary.

Right? So, I merge these two and then save as an edge. So, I did it.

So now I have edge data. Here there is 11,000 observations, the edges. And to import the base map

we need to use st_bbox and edge which is the data set that we saved. We need to use the same coordinate system with edges. Inside the shapefile there’s information about all the coordinate systems

and the locations, right. And we need to define the boundary and to define the boundary we

need to use st_bbox. So, we do that and if we see what is inside, it gives the four values about
the boundary. To use the get_map function we need to change the column name as a left, bottom, right,

top, that isn't fixed. So, if you do that and then see what happened, xmin changed the left, ymin changed the bottom. So, we changed it. This is required if we want to use a get_map function.

And now I want to get the base map and save it as a Glasgow map. So, this is what I’m doing.

Now I have a base map. You cannot see, because I didn't command the print, right. So, let’s do it here.

ggmap is how you print the map, so I said print the Glasgow base map here and then data is edges, this is the data, in data we have all the information of
edges and then total count because we merged them. Total distance, cycling distance. I want to

show the graph, each edges, but the thickness can change, the size can change depending on the

total travel distance, cycling distance, right. So, we see if it's thicker then it means there are

high levels of cycling activities. For the colour we use the Sienna1. You can change it
to the black, blue, red, whatever you like, right. Scale size, for the travel

total distance we need to define the breaks for the values, so we use 10, 50, 100, 150. I

will show you what it means, and range is 0.123. You can change it. This defines the thickness of the
breaks, depends on the breaks. And then we put the label, the title is total annual counting
distance per edges and size is kilometres. Let's see what happens if we run this one.
You see this nice graph. So, I will enlarge the plots. So, this is the kilometres our dependent
variable and this is the breaks we defined - 10, 50, 100, 150 - and that has a different
size, right. That's what we have here. And we use the sienna, this is a colour called Sienna1 and
that's the graph. We see the river and we see a lot of cycling activities
alongside the river and here. As you may remember in the previous session, we have this area and
this
area has a really nice cycling infrastructure, good cycling infrastructure and it totally makes
sense. It's not hilly, it's very flat. And see the city centre there are quite
good levels of activities. And then for whole real main roads there are some activities there.

So, this shows that, oh yeah, the data looks reasonable and that's what we expected and that's the
easiest way to see the travel patterns in your city, right. So that's what we did here. So, if we
want to change the key variable you can change it here. It depends on your research but then now
the main analysis that I'm going to do here is I want to examine the relationship
between weather conditions and cycling activities, right. The relationship between weather
conditions
and cycling activities. So first I need to process, I need to calculate total cycling distance

but for different purposes because covering cycling distance activities could be different from the non-commuting cycling patterns, right. So, we do that. How we do that? We want to calculate the total cycling distance per day because if we want to examine the relationship between weather conditions and the cycling activities we need to make our analysis unit as hourly or daily. But we decided to use daily here because weather changes by daily. So here I create total cycling data set but they use the Glasgow Strava data, which is this one, right. All the edge informations are there and also the activity Strava data, raw data. We now group
by day because we want to calculate total cycling distance per day. So that's what we are doing here.

and we use the summarise function to calculate total activities, means sum of total activity count multiplied by kilometres. So that's total cycling distance. For the commuting activities we use

commuting count, right. In the Strava data we see there was committing count if the users tick

this trip as commuting. And then kilometre again, the length of edges. For the non-commuting

activities then how can we calculate, we can use the total activity count minus commuting count.

That is the full non-commuting count, right, multiplied by kilometres. So then saved edge is

non-commuting activities. So, this total activity means total cycling distance per day,
right. So, we do that and let’s see what is inside the total cycling data.

See there is days 1 to 366. It’s omitted, it just shows the first six rows and total activities in day one there is only 886 kilometres. And for commuting none, because it’s a new year, right. No one will work there at the time. And non-commuting there are some people. So, we calculate total cycling distance for commuting and non-commuting by day. Here,

later, I want to use a nice graph and plot the date in a proper way, so I use
as date function. By using this variable, so total cycling data,

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

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

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.

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

01:11:53,360 --> 01:12:03,600
from day to date. And here, let’s do the sum. So, I calculated

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.

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
the total activities which means total cycling distance by daily and the mean is 2,500 kilometres for whole area but max is almost 10 times. That is weird. That's too large.

Then there could be some issue. So, what I'm doing here is I just want to see which date. So using the total cycling data, and if the total cycling data, the total activity is greater than 23,000 just plot them. That's what it is, what this command means. So, we run it and we have day 255 and that is September 11th. And I googled it, what happened in Glasgow on September 11th and there was an annual Glasgow to Edinburgh bike ride event on that day. So that is an exceptional date and for the analysis it's better to remove it. So here
I use the total cycling data, the same data, and filter so only select if the total activity is less than 23,000. So, if I do that.

now to the summary again, the max is 10k. You know, it's kind of better, much better, right. You can do more, if you are using the Strava data for your cities you can do more investigation.

But I think that's okay for this tutorial. So now I process the data, I check the data, whether there are errors or exceptional days. And then I want to see the trend of the data because each time series data is 1 to 366. So, I want to see the trend of the whole activities.

Here, I calculate the moving average. This is a good measure for showing the trend of the data.
If we use the raw data there will be a lot of spikes, so that is very hard to see.

But if you use the moving average it's nicer. Moving average means you average the whole past seven days and use that average as your value. So, I actually find this code from online, I mean there are a lot of R codes you can just google it if you don't know how to do it.

So, this is how we make a function to calculate moving average. I will briefly explain here the concept. So, this is the new variable that we are going to create, and this is the loop function.

So, let's assume that the i is the seventh of January, right.
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
total 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
total cycling distance for commuting purposes and non-commuting purposes. This is the variables.

and use this function and calculate moving average and save as a ma total, ma commuting

and ma non-commuting, right. And use mutate. Mutate is a function that when I want to make new variables.

So, I use the total cycling, again the final dataset we processed here and then save as

total cycling again. So, I don't want to make a different data set, I want to keep this original

one and adding more. So, when I do that, see what happens.

This command just shows me the first 10 observations and this one. They want total

activities, commuting activities, non-commuting activities, that's what we have. And date, that's
what we originally have, right. Now there is a new variable ma total ma commuting and I think

because the size is so small it doesn't really produce the other one but let me do it again.

Yeah, so the ma non-commuting, so this is the moving average. So, look here

there is no change until

five, fifth observation. Because if you look at this one if, for example, i equal 5 then this is 5 minus

6 because 6 is predefined it's -1. I cannot really calculate this number so when it becomes greater

than or equal to zero it calculates the moving average. So, this moving average is the average of
this. Seven [counts] actually six values. If you average this one there will

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

01:18:22,480 --> 01:18:28,800
and also, we have raw data, raw total cycling distance and total cycling distance for commuting

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

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

01:18:42,800 --> 01:18:51,360
is, I want to see the whole three information - total travel cycling distance, total commuting and

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

01:18:56,560 --> 01:19:03,680
because the current format has, for example, these three different columns. If you look
at the graph plot, the y, there's only one variable, right, so that's very hard. This format, we call it a wide format. Through the nice graph you need to transform this wide form to the long form. So how can we do that? It's very confusing, right, just hold on I will explain.

This is a function that we can change the format. So here, column 2:4 means I only use the column 2 total activities, commuting activities and non-commuting activities. 2, 3, 4.

And then use the names to activity type, create new variable activity type and put the value as a total distance. This is a new name of the variable, but I only use these three columns. So let's see what happens if I just run this code. Now see this one. Now it's the same but for one day,
January 1st, there are three rows, and each row has activity time, a type as a total commuting and non-committing. And the total distance here is the value of the original value 886 here, 0 here and 886 again here. So, I change the wide format to the long format. So, we have one key variable - total distance - and we know activity type, different activity type.

Does it make sense? And we use the ggplot to plot the trend. x is a date 1 to 366 and y is the total distance. And then we said oh let's print the points, each value,
the activity types we want to have a different shape and colours, right. And then this is the label

and again, scale bar the x axis is a date and I want to have this format -

year, month, and day. That's a nice format. And geom_line means let's connect all the points by line.

So, I will show you the final results, that's better to understand, for your understanding. So

I just use the raw data, not a moving average, and this is what it is. The total commuting activities, it has

a blue colour and square shape because we define here, right, different colour and different shape.

And it has this pattern. The non-commuting green and triangle and commuting activities

red and circle. So, we have a different shape for different activity type and also the colours.
See kind of seasonality impacts. There are low levels of cycling activities during the winter, here, but high activity levels during the spring, summer, and autumn.

There are some decreases because I think this is because of the holidays - there are not many students here. Also, people are taking their holiday. Again, this is not really easy to see the trend, so I use the moving average. It's the same command but I use now column six to eight because that's the moving average that we calculate 1, 2, 3, 4, 5, 6 so six, seven and eight, right. And the same thing, it's exactly the same code. So, if I do that
you see a much nicer trend, right. That is a reason why people use the moving average. You can

ignore this part because that's not really moving average. So, this is moving average. You can see the

seasonality impacts and also kind of variations because it's weekdays and weekends.

Also, weather can be the factors why there are such a big variation.

So now we have processed the Strava data. We know we have all the total cycling distance, cycling distance for commuting trips and non-commuting trips by day. So, we processed data

Now we need weather data, right, so then we can build a model to see the relationship between weather conditions and the cycling activities. So, there are two data sets that you can use. One is you can
get your data from your local weather stations, that could be more comprehensive.

But if you don’t have it you can use these two libraries to obtain the weather data for your city.

So first I want to calculate the length of day because, again, the length of day is very important for cycling activities and in Glasgow the length of the day changes significantly compared to winter and summer. So here I make sunlight data. First, I define the data frame the date. We need to ask them what date we need the data and then the location.

So, I use the total cycling date information that includes the 2016 January 1st to 2016 December
31st and then put as a date. So that's a data frame as in the data frame there's a date

variable, which is defined like this one and we put the latitude and longitude of our city area.

How can you do that? If you just google it, your city, Google will show you the latitudes and

longitudinal information, you choose information, so you can just type it. That is our data

and we use getSunlightTimes function to get the sunrise and sunset time, right. That's what

we do here. And we mutate length of day as the difference between the sunset and sunrise.

Then we can estimate the length of the day and we rename day date as a date no time because

sometimes date has its own function, so it can be confusing, but it's not really necessary. And then
we have a day; we want to have a day like 1 to 366 because it's easy to

use for merging other data set by using date no time. Date no time now is

January 1st, something like that. So, if we do that and let's look at the data.

and this is the date no time. It's 2016 January 1st and this format, we have latitude and

longitude information, sunrise information, sunset information. We have length of days, seven

hours it's been increased, and day 1 to 366, that's what we are doing here. So now we calculate

the length of the day for 2016. Now we need to get more detailed weather data, like precipitation,
station information for your cities. So here I also put the latitude and longitude for Glasgow

and then get the metadata inside the code. We see different stations, it takes time, so

these are the stations, weather stations in Glasgow. And you can use the code to select the

station. Here we select Prestwick, the weather data, weather station

in Prestwick Airport. The reason is it has a list missing value so we use that code, so we use

import NOAA function. We define the station and then we set each year to 2016

and only select date, wind speed, air temperature, precipitation, right. And the mutate date no time
again as a date, which means 1 to 366 and then hour of day. That's what we do here.

01:28:16,160 --> 01:28:20,400
So, we only select wind speed, air temperature and precipitation.

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,

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

01:28:38,800 --> 01:28:46,080
are doing here. So now we have all the weather data and rather than

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

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

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
weather daily, we calculate the mean temperature, max temperature based on the original data set.

I think that's straightforward.

Then we have weather condition data, and we have length of day data. The next one is we need to merge these two data sets so we can have full weather condition data. So how can I do that?

I have weather daily data, which is the final data set for weather, and sunlight data where I calculate the length of the day and both of the data sets has a day 1 to 366.

So, we use that variable to merge these two data sets. And now I have a final weather data set.
It says all the mean temperature, max temperature, precipitation, min wind speed, max wind speed,

latitudes, longitudes, and length of day. So that is my whole weather data.

Now what I need to do is merge this weather data with cycling data. In the total cycling data set

we already calculated this total cycling distance for commuting and non-commuting and also overall

total. We arrange it by day because we want to order the

data set by day. And then use inner join, same thing, we merge weather data and this Strava data

and then merge them. But before I do that, I want to make sure we have for the weather data

we have 362 days because there are some missing values, right. So, we do that and
then for cycling we have all the information. We have total activity, which represents the total cycling distance per day and the commuting cycling distance and non-commuting cycling distance.

right. And all the weather condition data, that's what we have. So that is the final stage. We have all the data processed, now we want to see the key dependent variable, how they are distributed.

So here I plot the key dependent variables. Total activities, again, total cycling distance, total cycling distance for commuting and non-commuting.

They are skewed, that's general, right. So, if we want to use some kind of regression model it's
better to make as a normal. So, we took here the square root, we take the square root transformation

of the key dependent variable and then see how it looks. Yeah, this one and this one

is much better. For the commuting still it's not really ideal, so you need

more investigation if you want to conduct a proper analysis.

But for this tutorial let's just go with it. So here, although this is the time series data

so the observations could be correlated, we assume that they are independent. So let's

just run the linear regression model, which means that we

assume that all observations are independent. And this is the square root of the total activities,
is our dependent variable and this is all our weather condition variables, right. We do that

and print the result. We see four weather conditions have very significant relationship with level of total cycling activities, cycling distance.

So, precipitation has a negative relationship. It means if the level of precipitation increases, more rain, the activity level will decrease, that's what it means. Max temperature increases the level of cycling activities. It makes sense because in Glasgow max temperature even in summer is not really that high. Max wind, if the wind speed is high there are fewer cycling activities. That totally makes sense. Length of day, if the length of
day increases, the total cycling distance increases and that also makes sense. These are all consistent.

with previous studies, right. I want to check the model result and then see the residual path from the model to check the model assumptions because linear regression model,

there are several assumptions. And it looks ok, actually, there are not many clear patterns and

then normal ggplot looks ok. However, again, there could be an auto correlation issue and then to

test auto correlation between observation we did a Durbin Watson Test. And then see there are

several lags that have a p-value less than 0.05, which means there are auto correlation.
issues. In that case you need to use time series data. So here I used auto arima function to use the time series models and then to do that I need a library forecast.

And this is the result. You don't have to worry about this other extra coefficient that's about the time series coefficient. But this one, if you can compare the estimate this one with the previous one. And what we found is very consistent. Although there are some differences in terms of magnitudes, the level of significance and also the signs are very consistent so we can conclude that yes there are significant relationships between weather conditions and cycling activities, total cycling activities. If you want to examine the commuting
cycling distance and non-commuting cycling distance, you can just change this variable, right, and that's the same. I checked the assumption of the time series model, but I still see some of the problems.

This is beyond this tutorial, so I don't want to talk about the models but as a researcher you may want to try other approaches to fix the auto correlation issues. So sorry for the long tutorial. I think it's a little bit hard to explain without your reactions but I hope you can get something from my tutorial. Again, this code will be available based on your request. So, if you need this code and data please do apply through the UBDC website. Thank you
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
the city of Glasgow I think is it Jin? [Jin Hong] Yes, that's the location that we use here. The whole data is a Glasgow one, so here, this area.

The data you will get for this tutorial is the Glasgow area.

[Muir Houston] Ok and just to remind everybody, as Jin has said, and I've posted the links to the data catalogue on the UBDC which gives information about gaining access to the data.

And I've also put the link there for the free GIS software QGIS and the link to R which is also free and open-source code. So, I'd just like to thank everyone for attending and thank Jin for
his presentation and workshop. And the recording of this will be on the UBDC website, but we will need to make sure it's ok for accessibility given new regulations about accessibility of online content. So once again, thanks very much for coming and keep an eye on the UBDC. We've got another three of these data dives over the next month.

So please, if you're interested, sign up and register for these and hopefully we'll see some of you at these other sessions. So once again, thanks everyone. [Jin Hong] Thank you very much, bye. [Muir Houston] Bye.