Exploring and analysing open transport data (with a focus on SCOOT data and bikeshare data)

Transcript from webinar video recording

1
00:00:01,643 --> 00:00:03,353
This is about, as I've said,

2
00:00:03,453 --> 00:00:08,862
this is about exploring and analysing the open transport data sets

3
00:00:09,700 --> 00:00:14,119
that are easily available to us these days.

4
00:00:14,338 --> 00:00:18,469
And, of course, it's not going to be exhaustive.

5
00:00:20,678 --> 00:00:23,079
What I'm trying to do is give some examples
so that we know where to look when we would like to find out more about our environment and the transport issues around us.

So, that's the goal.

It's not like I can give an exhaustive list of all these different data sets. That's not the case. But I try to introduce some in a way such that it becomes more accessible.
So, okay, let's get started.

And then, the overview of the session.

I plan to not use up the whole two hours.

I find that to be a bit too much for both the audience and myself, of course.

So, the first part,

I will try to keep to time and finish before 10:45.

And we will focus on this presentation
where I'll try to introduce some of the sources of open data

and then, how do we obtain them

and what do they look like in general.

And then, very briefly, I will go through a few applications that I came across in these days.

I know, we all live in a pretty strange time these days with COVID and everything.

For these applications, I'll focus on the later two,
which are about COVID and mobility

and then this is actually demonstrated,

de demonstrated by a blog post

on the website of the Urban Big Data Centre

written by our colleague.

And then, the other one would be

a paper that is, I think, quite close to publication

and it will be about the shared bikes in Glasgow.
And then, the trip data that we use and how from there we turned it into a paper.

So, that's what we are going to do.

Then, after a short break, we will go into the practical session.

And then, if you have asked to do, you are welcome to follow along.

If you don't, it's still fine.

I will share the script and the presentation afterwards.

So, okay, let's get started.
A little bit about me, not too much.

I'm currently a research associate in transport analytics at the Urban Big Data Centre of University of Glasgow.

So, during my PhD and after my PhD, what I've been interested in is mainly transport taxes such as road pricing, and then optimal public transport supply,
and also accessibility, which I’m currently working on.

And a few publications that I had with my colleagues are also about these issues, mainly.

The first one is about rural areas.

Our goal was to evaluate whether it is a good idea to have these kinds of very expensive rail systems in rural areas with very low-density populations.

And then, sometimes replacing it with buses is actually a better idea.
And then, the case study was from Sweden because one of our co-authors is from Sweden so we get access to the data.

And then, the other one, which is also using Swedish data,

we evaluated whether buses and cyclists, who are hindering who.

It’s very difficult to say because
if you imagine that,

I think some drivers are pretty annoyed by cyclists.

But then, in the case of buses

in a city area, like the city centre of Stockholm,

with a bus line, they have to go to the bus stops

and then stop at the bus stops.

And what happened is that

the cycle lane suddenly disappeared,
so the cyclists would have to move around, if you see what I mean.

They would have to switch to another lane.

Then they would have to interfere with the other traffic in the outer lanes.

So, that's the idea.

And then we tried to evaluate this interaction.

And then, in this paper, what we do is

we usually use a more aggregate model.
So, the key as to whether these models work well or not would be to evaluate the interactions between different modes.

So, that's usually the key.

How would one mode...

Like, okay, what if there is road pricing?

How will the ridership of public transport react?

So, that's the key idea.
The interaction between modes.

Why do I talk so much about this?

Because it is actually related to the paper that I'm going to talk about a bit later,

about the interaction between the subway system and the shared bike system in Glasgow.

So, okay, a bit about open data.

This is just a very brief introduction to it.
In the title of this session,

I actually wrote open data, I think.

But then, you need to be really careful that not all of these are Open Government License where you can pretty much, roughly speaking, you can reuse them,

you can basically redistribute them most of the time.

But then, you have to be really careful with the licensing terms
when it comes to some of the data that does not necessarily have an open licence.

So, yeah, you can see that many of them would be Open Government Licence later on.

So, I think that many of these data sources that we are going to talk about, it's usually quite straightforward in that we know what they are.
These tools, some are very straightforward,

like Department for Transport,

Transport for London, they have a very, very good page

with many different data sets.

And then, for some of the city council portals,

some of them I’m also going to

talk more about in the second half of the session,

sometimes you might need to register
to get a key or a login name to use them.

And it’s not exactly that easy because you might not get it in time when you try to use the data.

And sometimes the site might not be working very well.

So, let’s see how we get on with it today.

I believe that we included a link for one of the APIs for the Glasgow City Council data.

But then I found out that link might not work.
So, let's see.

Another very important source, of course, is the Urban Big Data Centre collection of data that if you take a look, very briefly.

And you can see that we acquire a pretty good collection of data,

such as this CCTV data.
where they process the image and the video

such that you could, if you specified the location,

you could get a count of the pedestrians or cyclists

or vehicles and so on.

And some of the very new stuff, other than that,

related to transport is this Huq data set,

which is mobile phone data
and very comprehensive and pretty detailed

00:09:30,879 --> 00:09:35,118
whenever the users use a certain app.

00:09:35,373 --> 00:09:38,124
Not a certain app. It's a collection of apps

00:09:38,224 --> 00:09:42,532
that the company works with

00:09:42,632 --> 00:09:46,209
so that we could have the location of

00:09:46,309 --> 00:09:48,767
the users.

00:09:50,265 --> 00:09:52,444
So, this could be pretty useful

00:09:52,762 --> 00:09:55,701
in terms of mobility research.
And you could go through this when you have time and applied.

Many of them are not open licence but then, of course, you can apply to use them from the Urban Big Data Centre and then be careful with the licencing part and see what you can do with them.

And also, one thing that I found is
At first, I didn't realise.

I know that GitHub exists

but then I found out that now,

sometimes when you browse through GitHub,

you find different packages and different scripts.

And that's led to some surprising discoveries

with data sets.
That's my opinion.

So, you might look through that later.

And for the map data,

Google Maps, of course, they have a lot of different data that you could use,

like distance matrices,

you could specify the route
and you can calculate the travel time

177
00:11:02,421 --> 00:11:05,259
and report it in pretty much real-time.

178
00:11:05,639 --> 00:11:08,408
But then, you will have to be very careful

179
00:11:08,508 --> 00:11:10,986
because you have to enter your credit card

180
00:11:11,524 --> 00:11:14,593
in order to use the API

181
00:11:14,693 --> 00:11:17,817
and you would not want to request too much

182
00:11:17,917 --> 00:11:21,135
and end up having to pay a whole lot for that.

183
00:11:22,025 --> 00:11:24,163
And also, OpenStreetMap.
That's also a very good source for things like points of interest and so on.

And if you were from academia or you were a student, if I remember correctly, you can also register to use Digimap.

So, these kinds of data sources, they could be real-time or historical.

Historical is simpler most of the time
if you know it's historical.

Of course, they would just dump the data on the site and then you can just click download and you get it.

It's usually more straightforward.

But then, for real-time data, usually it's more complicated because it gets updated a lot.

Like, every minute or every five minutes.
And if you want to get it

or research your purpose,

then you will have to make use of

some simple tools in order to

just get this data over time.

And how do we get it?

Of course, yeah, when I was saying about historical data,
most of the time you can download them.

And then, for the real-time ones,

most of the time you will need to make use of API

or, if you are familiar with it,

you can do some web scraping.

So, in the second half of the session,
to get access to some of this real-time data,

like every five minutes,

what is the count of traffic in certain locations.

Cars. We briefly talked about this.

The traffic count.

We have it.

The reasons that I put Brussels here is
because I planned to use

the Glasgow traffic counts

and the data sets as a demonstration,

but then I found out that

you may not be able to get an API key in time.

So, I think that, to avoid that issue,

to avoid registration,

we could try the Brussels one later
because it doesn’t require a key.

You can just request it using two lines in R

and then you get the count.

Okay. And also,

this is also real-time, obviously.

You want to know about parking availability
and you want to know it,

239
00:13:58,906 --> 00:14:02,717
to be updated in real-time.

240
00:14:03,697 --> 00:14:07,575
Sometimes, traffic events might also be interesting

241
00:14:07,773 --> 00:14:13,356
and you might want to know what happened in a certain part of

242
00:14:13,456 --> 00:14:16,284
the motorway, what happened with that part.

243
00:14:16,853 --> 00:14:17,967
If there were accidents

244
00:14:18,067 --> 00:14:20,568
or some sort of closure

245
00:14:21,067 --> 00:14:22,725
that we might want to know about.
And then, the next thing would be

some Variable Message System.

That's just the message boards you can see

when you are on the motorway.

When we try to look at them in the second half of the session,

then we could also show some of

these messages, what they say.

Like, over time, during COVID time,
they ask you to stay safe and stay home and so on.

And then, one thing that I was...

There's actually a picture of the Variable Message sign here.

You can see you can specify the location

and then you get a message about the roadworks at a certain point.

And another thing that I have been quite interested in are the potholes.

And yeah, I find that quite interesting.

FixMyStreet is actually not only about potholes.
But then, I actually laughed a bit when I saw this one.

You can see that people are reporting potholes because they are so pissed off.

And they report these kinds of dangerous potholes.

And they take pictures of them.

And then, if you read it carefully, if you try to read this,
you can even see that update.

270
00:15:44,259 --> 00:15:47,547
I thought that it would be from the city council or something, but no.

271
00:15:47,647 --> 00:15:49,780
It’s actually from another user,

272
00:15:49,880 --> 00:15:53,308
saying that there’s not a chance they’ll repair it or deal with it.

273
00:15:53,616 --> 00:15:55,695
So, it has been quite interesting.

274
00:15:55,795 --> 00:15:58,642
So, if you were looking at a road network,

275
00:15:58,742 --> 00:16:02,806
I suppose this information about

276
00:16:02,906 --> 00:16:05,373
the quality of the road would also be
Quite useful.

So, of course, accident reports are also part of the picture.

And, yeah.

Then we move on to another mode with bikes and pedestrians.

Then we have this, again, from UBDC.
Where you have this API that you can request the count of pedestrians and other vehicles or cyclists over a certain period of time.

You could also go over with cycling infrastructure.

These are links. But I don’t think I will show every one of them, but I will share it later.

And then, one focus that we will be talking about in the second half of this session
and also, it would be related to one of my papers is the use of shared bike data.

And there are usually, if you go on a site of a shared bike company about their data sets,

there would be two different links.

Usually, they have this data dump
where they have a lot of files

00:17:30,859 --> 00:17:33,907

about the details of every trip.

00:17:34,007 --> 00:17:37,276

When a bike is unlocked and when it is returned.

00:17:37,595 --> 00:17:38,723

And so on.

00:17:40,500 --> 00:17:43,879

I find this to be the most useful one

00:17:43,979 --> 00:17:47,008

and then we make use of this

00:17:47,108 --> 00:17:48,726

in one of our papers.

00:17:49,093 --> 00:17:52,113

And the API for shared bike is
actually quite easy to use.

And if you are interested in the availability of these kinds of bike,

or how these bikes are arranged

or moved by the company,

or how the demand is different,

then this is also quite useful.

But then, from this API, as you can see later,

when we actually tried it, we could see the location,
the co-ordinates of the bike station,

and also how many bikes are there

or how many spaces are there,

instead of other trip information

from the file that you can download as a data dump.

I find that it's also an interesting point

if you are interested in bike sharing.

If we look at this from a different point of view.
Like, okay, this is in Edinburgh.

You can see that you can download the trip data.

And this is what I was referring to as a data dump.

You can see you have all this information.

Like stop and start station,

description,

co-ordinates, and the time, and so on.
So, you have very detailed information for each trip.

So, if you go on this data,

that’s what you get, and then you also get

real-time data about the station’s availability

that I was referring to.

But then, if you click on some of the US bike share data,

I think it's also quite interesting because

it looks very similar.
But then, what strikes me is that they usually provide more information.

I guess this is partly due to GDPR.

That's my guess. I haven't really looked into it.

But then, for the ones outside Europe, you usually get something more detailed.

Such as the user type, even the birth year
and the gender of the user.

So, it depends on what you need in your research.

That's a point that is worth noting.

So, public transport.

It's actually, for me, a bit painful because it's not very easy to obtain data.

It's mostly not shared amongst the public.

And most of the time, you will need to obtain them directly from the company.
And then, as you can see,

if they are not in a working relationship with

you as a researcher

or if you are not a consultant working for them,

then it's not very easy to get sensitive information.

But then, one thing that is open, at least, is

the timetable or even, most of the time,
they will share the real-time location of their vehicles.

363
00:21:21,223 --> 00:21:23,001
So, that's also possible.

364
00:21:24,284 --> 00:21:27,291
And for the prices of their fares,

365
00:21:27,589 --> 00:21:30,257
it's a bit easier because

366
00:21:30,357 --> 00:21:34,595
you can basically scrape them directly from the web.

367
00:21:34,976 --> 00:21:38,914
Or in the case of rail here,

368
00:21:39,014 --> 00:21:40,202
Rail Delivery Group,

369
00:21:40,302 --> 00:21:44,790
you can also register and get this information from them.
So, I think this is the most constrained area in terms of getting data that is available to the public.

We don’t know that much about them. Especially in this country, with the market structure of the operators.

That’s pretty unfortunate because it’s a big missing piece of the puzzle, I would say.

Someone said freight data is even more constrained.
Yeah. Yeah, that's even more unfortunate.

Okay. So, I'm just thinking that, okay, with Glasgow, with these kinds of data, if we put them all together, we have different modes. And, of course, it's not like a complete picture. But we have a much better picture if we gather them together and try to build
models, including different modes.

And for the mobile phone data,

again, I highlighted it because

UBDC, they are having this mobile phone data set.

So, yeah, perhaps if you are interested,

just check it out and apply to use some.

I will move on to some of the applications.
One thing that I've found about

this type of data available to the public is

this I found during times of lockdown.

I read an article, I think it was in the Financial Times,

or The Economist.

Both of them did something pretty similar.

They tried to compare the hotspots
during lockdown and so on.
It was very early in the lockdown,

like, last year, in March or so.

So, there is actually this...

When you search for a place on Google,

for example, this is Glasgow Botanic Gardens,

which is like five minutes from where I am now,

what you will get is you get some popular times

and you see what times at this place are
usually the busier times.

And then, what I did is that I also found something pretty interesting,

which is a GitHub repository,

which I didn't write,

that enabled us to download this data

in real-time so we get an idea about

how busy a certain location is.

As long as you can get the location ID,
the details are on the GitHub.

It's the link here.

I'll just press it.

It's something that looks like this.

So, as long as you can get...

Again, you need this API key to set it up.
But then, what happens is that,

00:25:01,979 --> 00:25:04,500
if you get the ID of the park,

00:25:05,662 --> 00:25:11,354
then you could do something that I was trying.

00:25:11,943 --> 00:25:14,724
I got a list of the IDs of some parks

00:25:14,824 --> 00:25:20,334
and I tried to compare them to see whether they became more popular

00:25:20,434 --> 00:25:22,400
during lockdown or not.

00:25:22,901 --> 00:25:25,692
And I find that this is actually quite interesting

00:25:26,083 --> 00:25:28,444
because you can see that, on the left,
there is this one with...

The blue bars actually represent,

during normal times,

how popular the park is.

And then, the two lines represent the lockdown.

So, you see that some parks are actually less popular,

some of them are more popular

during the lockdown.
And then, when I look at...

I don't know Glasgow that well.

Actually, I wasn't here for that long.

But then, the interesting thing is that...

Where was my file?

But then... Never mind.

I was just talking about that.

I could do that without showing that, actually.
So, the parks, about the parks, which have this pattern where they become less popular during the lockdown.

They are actually the city centre's smaller parks.

One of the ones that is becoming more popular is actually in the peripheral area of the city.

So, I found that to be quite interesting.
Although, it seems to be a pretty trivial observation

but then, yeah, you can see all sorts of things when you try to

carry out this type of analysis.

So, that’s the first one.

And then, this one we will also get into more detail with a bit later,

although not exactly using the API of

the SCOOT data in Glasgow,

in the second half of the session
where I try to...

If you read the map carefully,

you can see the green icons.

This one, it says,

"Check your mirrors when moving between lanes".

So, this is actually a map

showing all the...

If you read it on the legend.
This one shows the Variable Message System.

The messages.

Here, you also have, "Look once, look twice", "Think bike", and so on.

And then, in the middle,

the orange ones are actually about parking availability.

And then, the blue ones are the count locations of traffic.
So, these types of things,

you could see that if we get hold of

the real-time data,

we can easily visualise them and download this data

for research purposes.

And the third one is actually about

the paper I'm going to discuss
and I'm trying to introduce it very briefly

\[00:28:28,605 \rightarrow 00:28:33,733\]
so that we could go through some of the steps later

\[00:28:33,833 \rightarrow 00:28:36,276\]
in the second part of the session.

\[00:28:37,317 \rightarrow 00:28:40,265\]
So, this paper is about examining

\[00:28:40,634 \rightarrow 00:28:43,851\]
the effects of a temporary subway closure

\[00:28:44,506 \rightarrow 00:28:47,428\]
on cycling in Glasgow.

\[00:28:48,548 \rightarrow 00:28:53,046\]
So, we used this bike sharing data from

\[00:28:53,231 \rightarrow 00:28:56,545\]
Nextbike which operates in Glasgow,
and then what we want to evaluate is,

there was a temporary closure of over one month in 2016

for the whole subway system.

So, that's quite amazing to me

because I've rarely heard of anything like that.

You close a whole system for a whole month
for renovation and then upgrading and so on.

And then, since the system is quite small

and then the distances between the stations,

the subway stations, are actually pretty short.

So, I was thinking that, maybe,

people will replace these subway trips

with cycling trips.

So, that's the idea.
And then, that's why we evaluated using the bike sharing data.

So, again, this is the same objective as I was saying about some of my previous work.

To study how different modes interact when there is this disruption of the public transport mode, subway in this case.
How will people react in terms of their cycling behaviour?

That's the idea.

So, very briefly then,

we narrowed down the study period to here

because you have this red segment

which is the actual suspension from July to some time in August.

And then, these phases are just the introduction of the bike stations,

the number of bike stations

that Nextbike are having during that time.

I excluded some of it because the growing number of bike stations might have affected the results.

So, that's the idea.

We study a part within the month.
the system is actually pretty small.

So, this is just to show that.

These stations are just...

If you're not familiar with the system,

I found it very amusing when I first came to Glasgow

because it's just a circle

and you go clockwise or you go anticlockwise.

That's it.
You can never get lost.

And it's very, very short.

So, that is why I think that it is possible that it can be substituted with bike rides.

So, what is this about?

It's not super clear.

But then, you can see the orange ones are the subway stations
and then these little dots are the bike stations.

So, the idea is that, of course, we try to compare the time during the suspension and after the suspension to before the suspension so you can see how behaviour changed.
But then, that might be problematic because we do not make use of most of our information.

In terms of bike rides, we know the starting stations, we know the ending stations, so we have more information than that.

So, what we have done is that we draw this buffer area
where it's like a catchment area,

where we think that, okay,

if these bike stations are inside this treatment area,

this catchment area,

then we will call these bike stations

the treatment stations.

The other ones are the control stations.
Like, these ones that are very far away from the subway system,

00:33:15,836 → 00:33:17,134
the subway stations.

00:33:17,522 → 00:33:20,660
So, the idea is that

00:33:20,760 → 00:33:25,046
these stations, the trips of them are affected by

00:33:25,146 → 00:33:27,954
the subway suspension, whilst these are not.

00:33:28,107 → 00:33:30,865
So, we get another comparison,

00:33:31,410 → 00:33:33,630
instead of only over time.

00:33:33,730 → 00:33:36,998
We compare the suspension and without the suspension.
So, that's the idea.

So, yeah, of course, it relates to some of the existing studies.

But then, again, I will not go into detail about this.

It's just that, if you are already familiar with the terms, then you will know this is only about the interactions between the different modes.

So, to put it very simply,
it's just that now we are suspending the service of the subway.

That means, you can think about it as an increase in the full cost of taking a subway trip.

And then, with this change, how would it affect the quantity of trips in bike sharing.

So, that's the idea.
And then, if we put together these types of measures,

then we can compare these with different studies.

That's the idea.

And then, sometimes, if you go into the literature,

it's not a new thing that people keep wanting to

find out about the substitutability of, or
complementarity of trips.

Like, they study Uber as a complement or as a substitute of other trips.

And also, it is possible when you look at it about rail or bus and so on.

Because if you see bike sharing as a substitute, that means it replaces the trip.
But then, if it is as a complement,

then it is like the second case here

where it is a first-mile last-mile facilitator,

which means you ride the bike to, let’s say, the rail station

and then you take a rail trip.

So, that’s the idea.

And then, of course,

this may look a bit intimidating
if you are not used to this,

but then, the idea is, as I've said,

that you're making two comparisons.

One is between the time periods

when the suspension is there,

and since we would want to see whether people are coming back to,

again, the...

Let's say, with the suspension,
they want to ride the bikes
and afterwards, do they still want to keep on riding bikes?
Or when the subway is back, they will switch right back to
riding on the subway,
So, that's the idea. So, that's why there is
the post-suspension period also.
And then, another comparison would be
the treatment station and the control station,
the bike stations that I was talking about.

So, when you do the analysis,

you create different variables and so on.

I won't go into details.

If you followed the study, then you would know that

after some basic cleaning of the data,

you would want to cut some of the trips that are

perhaps way too long to be covered
because we are talking about replacing subway trips

so it won't make sense if they are riding bikes for five hours.

Then I can't really claim that it is a commuting trip.

So, that's why some of the trips are removed

and you can see that the percentage is not that much

if you draw the cut-off

around 150 minutes or something like that.

That's pretty useful in looking at
the whole picture

and deciding where you want to make the cut-off.

And then, also you have the...

This is a diagram for the justification of using this type of model where I have the treatment and the control and so on.

You can see that,
visually, the treatment group and the control group do behave quite differently during the suspension period.

So, I think that’s quite strong visual evidence of that.

So, the hypothesis, to say it simply is that since Glasgow subway station is in a dense city area, we would expect that these two modes, the bike sharing mode and the subway, would be substitutes.
the coefficient would be positive,

the representing coefficient of interest.

So, yeah, I wouldn’t go back to the notation

but then the findings show us that, actually,

despite the system being pretty small,

we really get more incoming and more outgoing bike trips

when compared to the control group

and also when compared to the other times that
were without a suspension.

So, that kind of

strengthens our argument for

our initial anticipation for the results.

And then, although the number doesn't look too large,

it looks like, okay, it's not a large number of trips,

but then the system, both the subway system

and the bike sharing system are not that big
so it already represents around a 20% increase

when you compare it.

So, it's actually quite big,

But then, of course, some of them,

after the resumption of the subway service,

they disappear again and they go back to

riding the subway.
So, after this,

then we will see that,

with these findings,

if we converted...

If you remember that earlier

I brought up this term, diversion factor,

it's just that when we have one fewer subway trip,

like, people are shifting away from the subway,
how many of them ended up riding a shared bike?

It would be very, very small.

Less than 0.05, which is the suggested value in some of the literature.

But then, of course, this could be due to many reasons,

that we found a much smaller value than this.

So, I think this will end the first half and then the second half, we will get more hands on with
trying to obtain some of the data

and do something with it.

In the second half.

And then, I will also try to show

how we start from there

and then what could be done potentially

with the data set to create something that is

pretty similar to what I have just shown you.
Although, I couldn't really use the Nextbike trip data because it's obtained from the company and it's not on their website right now.

But then, we would be using something that is freely available in the session.

So, this is it.
And yeah, we will come back to the second half.

719
00:41:26,957 --> 00:41:29,975
But then, do I have to...

720
00:41:30,831 --> 00:41:32,590
Shall I answer something?

721
00:41:32,960 --> 00:41:34,736
Some questions.

722
00:41:35,826 --> 00:41:38,603
So, thank you, thank you for your talk.

723
00:41:38,975 --> 00:41:41,344
I think the very first question,

724
00:41:41,444 --> 00:41:44,786
which is not a question, rather a comment, was

725
00:41:44,886 --> 00:41:50,484
about Norway's system for public transport
public information portal.

- It's compulsory, it's cool.
- Which I think is very...

Yeah, it's...

So, in Norway,
there's compulsory sharing of
the information of every stop and timetable for public transport.

That's pretty cool.

I think that, if I understand correctly,
we are also going towards that direction.

But we are not there yet.

From what I understand.

Yeah.

I would encourage everybody to look

in the chat room, there is a useful link

pointing to English development pages,

if anybody is interested.

So, public information portal for Norway,
which is based on open data.

Right. Let's move to the other questions.

A question from Alan.

How are you accounting for the other counterfactuals which have affected the treatment group?

I think that we... Yeah, there are, of course, there are a lot of things that we couldn't really take into account.
Then, the model we used.

First thing, it was,

we tried to take in the differences of stations

in terms of using the fixed effects,

that's the first thing,

for different stations.

And also,
we included the weather data

and then some other variables.

And at the same time, of course it won't be perfect

but we tried to make sure that,

at the same time, there were no particular things that were particularly affecting the treatment group only.

So, I think that's probably a simplification

but that was the best that we could do
from that simple model and the information that we had.

So, it was a panel fixed effect and also, the other co-variants

and trying to take out the period, as I have shown,

the diagram with different phases.

So, we tried to identify, we tried to cut out the time period
where there were a lot of things happening with stations

because there had been an increase in the number of stations.

So, that was what we were trying to do.

I hope that answers the question, partly at least.

Okay.

Sharing the presentation as a PDF?

Yes, I will.

Actually, the scripts that we are going to use
and the presentation will be shared.

But I’m not sure if it will be on my personal GitHub or the Urban Big Data Centre’s GitHub.

I will try to find it out.

And then, when we share this recording, it will also be available.

Can I just quickly jump in with a comment?

In a previous session, I think,
every registered participant received a link to a recording

and supporting materials at the very end.

- So, I think it's...
- Oh, I see.

- That is cool.
- This time as well.

That is cool. I haven't uploaded it yet

but I will tidy it up after today.

So, just to answer the question.
It will be shared in some form.

- Yeah.
- To registered participants.

Thank you.

If you want to follow along for some of these apps,

I'm trying to...

Do you see my screen with RStudio?

Yes.

Brilliant.
So, I think that the best way is that...

Some of these scripts, of course, because you don’t have the script with you,

I haven’t shared it yet.

So, it would be unreasonable to ask you to

follow it, unless you type in a super speedy way.

But then, I think if you want to follow along,

for some of the steps, I will share
the particular command in the chat box

so that we could try running it for some parts.

But then, this first part, I would not expect that

because it's a bit long and you can try it later.

I guess. Because this part is

actually using the API for Glasgow City Council

on the information that they share,
as you can see here very briefly.

00:46:57,397 --> 00:46:59,514
It would be about

00:47:03,637 --> 00:47:06,944
the movement, the traffic count,

00:47:07,478 --> 00:47:08,726
the car parks,

00:47:08,826 --> 00:47:11,814
the events that might be happening on the highways,

00:47:11,914 --> 00:47:13,271
the locations,

00:47:13,371 --> 00:47:18,628
and the Variable Message System that I was talking about.

00:47:18,914 --> 00:47:22,312
So, you will need an API key
and then, we provided a link to access that,

to sign up for that

for this session.

But then, I believe the website was dead for quite some time.

I don't know if it was my problem or not.

If any of you could actually get a key or could actually sign up.

So, it doesn't really matter. I will still share the script

and you could try it later, after you are able to get an account.
But we could see what it looks like.

So, the very first thing is that there would be some libraries that we might need most of the time.

Some libraries like jsonlite and then something for tidying the tables.

And then, one particular library I would be using would be Leaflet, just to show some locations information and then it looks pretty nice.
The HTML tools are also related to that.

And then, yeah, you could sign up later.

You would be able to, if you got access to the site,

you would be able to sign up for a standard account,

or something like, start an account.

And then, if you would like a lot, of course,

over time,
like, making a request every minute or whatever

00:49:00,814 --> 00:49:02,395
that you would like,

00:49:02,612 --> 00:49:04,510
then you would need a business account

00:49:04,610 --> 00:49:07,397
but then, yeah, you could do that when you sign up.

00:49:08,167 --> 00:49:12,235
And then, there are a few URLs.

00:49:12,335 --> 00:49:17,251
And the key here is actually just entering the key.

00:49:18,066 --> 00:49:20,843
I didn't show my key, I have already entered it.

00:49:22,129 --> 00:49:27,026
So, this URL is provided from their website because
it's, from what I know,

it might not be accessible right now.

So, I would just not show it.

But then, the flow is a bit like this.

You get the key, you get the URL,

and then you request that using this command

by placing the key if it is necessary.

And then, the next part would be
more problematic because you will want to see,

after what you have requested,

how it looks and whether you can just use it directly

for your, like, will it look like a very nice data set

that you can use right away.

So, yeah, it works, which is a good sign.

So, from this request,

we can see what we actually have.
And then, if you look at it, then you can see...

Okay, this doesn’t look like what I was expecting.

I was looking for traffic counts.

And then, if you open it,

there is some information about the whole schema.

And then, you continue with it.

Then you will start seeing the information
that you would want.

Because here, let's say this one is number one,

and then you will start seeing the level of the traffic.

Yeah, you see the traffic flow of that particular location.

So, you can see that the site information is here.

Each of them is a different site.

So, this is not tidy at all
and most of the time, when it comes to counts and locations,

the counter information, the data,

that you try to grab,

as far as I know, most of the time,

there are no simple commands that you can just use
to tidy up this type of data,
So, of course, I would think that it would take too much time for us to do this like this here.

But then, I guess the point would be that if you click it and you can see that there are different levels and layers. And then... Yeah, I will have to move this a little bit.
So, that was the request.

And then, we tidy it up.

I did some of the following steps to tidy it up.

And then, it looks a bit weird

but then, the point is that

I tried to get the movement information

from those different layers
and put it back into a data frame how I would like it to look.

911
00:52:55,964 --> 00:52:59,493
So, these are a bit annoying

912
00:52:59,593 --> 00:53:00,817
when you try to do it.

913
00:53:00,917 --> 00:53:04,455
But then, we will try another one later.

914
00:53:04,683 --> 00:53:08,144
But then, this is how it will look

915
00:53:08,244 --> 00:53:09,582
at the very beginning.

916
00:53:11,739 --> 00:53:15,359
So, we could see that this is just requested.

917
00:53:15,459 --> 00:53:18,237
So, if you click that open,
you will see that the time is just now, not far from now.

And then, if I run these,

then I will be able to get a data frame called movement count,

which is the camera

and the number of the count,

so that I could make use of it more easily.

And then, let's say if....

I will just try it with the parking data,
doing almost the same thing.

Requesting through the URL and then tidying it up.

Okay. I run it.

Then what I'm getting here is that...

Okay. We'll try to open it and then we will see that there are

the timestamps,

and then you can see the parking locations,
and also, whether there are enough spaces and how much of it is occupied.

So, this sort of information is what we could get from this particular URL.

It is a bit annoying because, I'm sorry, the bar keeps blocking me.
We can move it.

Yeah. Finally.

So, the parking co-ordinates,

they look like this.

And then, there's the traffic events.

And then, I could do similar things to it.

And also, the Variable Message boards.

And you can see that after you've requested it,
it's also a bit of a mess if you try.

This is the message board.

And then, this is the thing that we have requested

and we try to press into it.

And, we will see something, like,

the information, the status,

and then the message that is actually there.

So, what we have to do is
that we have to extract the useful information from this.

And then, that's what we get.

It would be after the steps.

Because these are actually just columns.

You go into one column and then you go into another column.

So, I didn't go into great detail about this.

And then, I renamed these columns so that it will look nicer.
And then, the data frame, if we tidy it up,

it will look like, okay, we have these five locations

and then these are the times.

And then, the message is just something like this.

"Check your mirrors when moving".

And so on.

Sometimes there is more interesting information
but not right now.

And, of course, these message boards also have their location.

And similarly, we could request this.

So, what I was doing with this is...

The next thing would be to combine these message boards with the location, because the message board data set and the location data set are
in two different data frames right now.

So, if we try to combine that, I have to rearrange a little bit because some of them, they have extra characters if you click into that particular name of the data frame. So, I remove some of them and I try to combine them.

And then, in the end, what I aim to do is nothing too complicated.
It’s just that I was trying to put together these data sets and then place them into a nice map. So, I will make use of the Leaflet package now and then I add different markers, such as I supplied the location and what kind of label I am going to use. And then I save it as an HTML file. So, this is for the parking map.
and so on.

And I try to choose different icons for that.

So, I think the most interesting one that I'm going to show is the one where I included the different markers on one map.

Like the one for the parking locations and the one for the count of traffic.
and also, the Variable Message Systems.

00:58:34,842 --> 00:58:38,812
And it will look something like this.

00:58:41,331 --> 00:58:42,969
So, you have something like,

00:58:43,337 --> 00:58:45,766
okay, the messages are showing here.

00:58:46,322 --> 00:58:50,961
And then, if you go around, if you move around.

00:58:51,790 --> 00:58:53,388
This is way too slow.

00:58:53,933 --> 00:58:58,172
You will see the count on the map

00:58:59,590 --> 00:59:03,827
along the way. Or when you hover over the parking ones.
I've got a flag. I couldn't find a parking logo.

And then, you will see that

I am showing the message whether there would be enough space.

So, I think that could be something that would be of interest if you want to combine them and show them all.

You find that too chaotic?

Then you could remove them.

You could overlay each one of them
or you could just remove these.

And this is done using the Leaflet package.

Then the next thing would be,

I think that we could try something with this one on Brussels' traffic count.

So, let me go to that one.

This API, because it's open without a key.

So, I think if you have R with you,
you could just try it.

Let me share this.

Where's the chat box?

Okay. There you go.

So, if you go to this.

This is the site.

Then you should be able to see that
this is actually more complicated

than the one that we had for the Glasgow City Council

because of what they supplied.

We mainly have the vehicle count for that

and this one, we actually have way more information.

You can read it in detail later,

but then the point is that

you can have this kind of live data
about the count and speed and even the occupancy of the road,

and the timestamps and how they get hold of these numbers.

So, if you look here, you get a link with the livestream counts.

And also, this one, if you click on this one,

you would be able to download a CSV with the locations of the devices.

So, that would be useful if you want to put them on a map

and if you want to know the exact location of these devices.
So, let's click on this.

And then you see,

when you place a live request, it will mostly look like this

for a lot of the cases you will look into.

First of all, you get the request date.

This is updated.

They are in Brussels, so they are one hour faster.

And, of course, it's close to this time.
And then, you get the device

and you get it for one minute.

And you get the count.

And then the speed and the occupancy

and all that.

It's not only one minute you get.

I think you get five minutes, fifteen minutes, and so on.
But then, the main problem would be,

of course, to get this and tidy this up

First things first, we just try some of this

and with the URL that we just had.

Which is this one.

You can either save this page and then import it to R

or you can copy this one.
I think I just shared that too.

Yeah, but I can do that again.

I think we can put this up, sorry.

Put this in R and try to import that.

Yeah, like, okay, we start from here.

URL. So, we enter this URL

and then we do this request.

If you have RStudio with you,
you could also try.

And then, if you run it,

then you will see that we have the whole request for Brussels.

This is the data that we are looking at

and you can see that's the device

and then the results involved with this device

and then you have this different type of data

that you get.
You have the count and so on.

There's none for this one.

Yeah,
Perhaps, again, another one.

Then you see this kind of information.

That's the type of thing you can extract.

Of course, then we will have to go through the annoying steps of requesting them.
I didn't include them here because I will put them,

I would not want to go over them here

and I will include them later when I share the script.

But then, this is the idea.

And then, as I've said,

the location of the counter is actually a CSV file that

you could download from the page that I just shared.

Then, afterwards, I think the next step would be
to combine the counter location

and also the count.

And then you can do different things,

like putting it on a map or something like that.

Of course, that's the map.

But then, I didn't go into a lot of detail.

Where is it?

This is the Brussels map
where you can see these are the different counter devices

that I plotted from the previous CSV file.

So, you can show more information,

like the count, after you have combined the data sets.

Again, I will share this information,

this script, later.

But then, it would be interesting
if you really wanted to try to tackle the area of your interest

1120
01:06:22,353 --> 01:06:23,408
for traffic count.

1121
01:06:23,508 --> 01:06:26,687
You will have to try to do that on your own.

1122
01:06:26,906 --> 01:06:30,485
Because I found out that, sometimes, for different cities

1123
01:06:30,585 --> 01:06:31,905
or different countries

1124
01:06:32,137 --> 01:06:36,209
the way they share or structure the scripts would be

1125
01:06:36,309 --> 01:06:40,237
a bit different, and it's not entirely clear to me

1126
01:06:41,087 --> 01:06:44,136
if there's a more uniform way of doing it.
But from what I've found also on GitHub and from other researchers,

they do it in this kind of way, one by one.

So, it's a bit annoying

but then you get what you want.

After you set it up, you can use it,

probably, for a longer time

so the effort will be worth it.

Actually, there's another example for Hull.
And if you check, this one also doesn’t...

It’s super small, sorry.

This one also doesn’t require an API key,

so you could also try to play with this one.

You can see that if you don’t want to deal with

all the messy files that I just showed you,

you can also download the CSV.

But then, you can see that it’s one location by location
so you have a ton of files if you want a lot of them.

There are quite a lot of them.

And then, if you scroll down,

then you can see that you go back to the original things.

scoodata.geojson file or .json file.

Let's see if we can access this.

Then you get this URL.
And then you almost go back to something very similar that

1151
01:08:37,215 --> 01:08:39,534
you will have to read.

1152
01:08:46,543 --> 01:08:50,718
Okay. This is the one that I'm referring to.

1153
01:08:50,818 --> 01:08:52,194
It's a bit slow, sorry.

1154
01:08:55,656 --> 01:08:57,091
Why is it not moving?

1155
01:09:00,344 --> 01:09:02,584
Yeah, you go back to something like this.

1156
01:09:02,836 --> 01:09:05,095
So, again, I would say that

1157
01:09:05,305 --> 01:09:08,223
this is going to take some time.
And, yeah, I will share it later, how to go fast through this because I don't think we have enough time to do it right away.

But then, one thing that I would like to mention is that these things are usually in real-time and it's not much use if you just get the data at this point in time.

So, how do you want to use it? I think that it would be quite interesting
if you want to try it.

You can connect through a PostgreSQL database

and then you save it every certain number of minutes

and then you build your database

will all this kind of information.

And then you can schedule your script

to be run in a certain time interval,

so that you can get this type of information.
So, yeah, this is what this part is doing.

And then, if we move on to some of the bike share data,

there are many lengths that some bike companies

that I have found over time.

And then, the thing about bike share is

that it's much nicer than what I've shown you about the traffic counters.

So, let's say for CitiBike.
Let’s go to the site of CitiBike.

01:10:49,542 --> 01:10:52,091
Why can’t I click on the link?

01:10:53,379 --> 01:10:54,748
It doesn’t really matter.

01:11:02,078 --> 01:11:04,316
CitiBike, here.

01:11:17,267 --> 01:11:21,242
I think I have shown some of this in the previous presentation,

01:11:21,342 --> 01:11:26,106
where you have all the trip data you can download as a data dump

01:11:26,206 --> 01:11:31,533
and then you have this kind of real-time data in this form.

01:11:32,275 --> 01:11:38,632
So, we try to obtain the real-time data here
on the script because...

About the trip data, I will briefly talk about it a bit later using the London shared bike.

So, again, we do the same thing.

We obtain the link.

We don't need an API key so it's relatively simple.

We get this.

So, the response of the request would be called Citibike.
And let's see what Citibike looks like.

It's not too bad. You have the station IDs, station names, and then the locations, the co-ordinates, and the status and so on.

So, this is relatively simple.

And then, if you try to arrange it a little bit from this Citibike data frame,
I called it the station list.

I arranged it a little bit.

Then you can see that we could obtain this data frame which is quite tidy,

if you try it like this.

I can also put this in the chat if you are trying.
You can just run this and then you can check the column links

1213
01:13:07,614 --> 01:13:10,018
and then you can see that this is quite neat.

1214
01:13:12,118 --> 01:13:13,753
Station, yeah.

1215
01:13:13,853 --> 01:13:14,967
Can check it.

1216
01:13:17,357 --> 01:13:21,628
You can see that you have the available docks

1217
01:13:21,728 --> 01:13:23,733
and also, latitude and longitude,

1218
01:13:25,562 --> 01:13:26,980
and the status.

1219
And then also you can get the timestamps.
Oh my god, this is from 2016.

But it's supposed to be much more up-to-date.

Most of these different companies, they work in very similar ways and it's quite straightforward.

And then, because they have this combined way of inputting bike data, usually it's not as complicated,
and then you could try on your way.

And then, with Nextbike, the bike company that's also operating shared biking in Glasgow,

we could also obtain that in a very similar way.

Like the station data and the status and the number of available bikes.

So, yeah, this could be done.

So, this brings us very briefly to

the last part.
Actually, this is a very simplified version of what I... I intend to share a very simplified version of the paper I was talking about earlier.

Which is something that I am trying to share.

Wait a second. It's a bit slow.

So, that was the analysis, the application that I was talking about a bit earlier in the first half of the session.
So, the workflow is a bit like this.

Of course, I don’t expect you to follow every step right now.

But then, of course, you import the libraries.

And then, because this is a data set

with the detailed trip data.

You can see it looks like this.

You have the start time, end time, duration,

and then also the station number.
And you can do all sorts of things,

including arranging these numbers in terms of stations.

So, the idea is that we could also...

Because when I was talking about a paper, you remember that we had this suspension period and so on,

so that we could introduce some dummy variables
to mark the period during the suspension

and after the suspension, and so on.

And also introduce the treatment group

and the control group for the different stations.

So, the workflow is something like this.

Where you introduce the buffer

and then you introduce the dummy variables

for the treatment and control stations,
the suspension period,

and also clean up the data about longer trips

that you might want to draw up

or trips that you do not want to include in your analysis,

such as having the same origin and destination.

We also include weather data.

So, it also took some time
to just put in the weather data by day

01:17:07,597 --> 01:17:12,114
and arranging them such that it could be combined with

01:17:12,214 --> 01:17:16,483
the panel data set and also the public holidays and so on.

01:17:17,621 --> 01:17:21,579
So, if we go back to the R Script.

01:17:22,057 --> 01:17:29,452
Of course, because I couldn't share that particular data set.

01:17:29,789 --> 01:17:33,647
So, that's why I think that we could do it like this.

01:17:34,084 --> 01:17:36,843
You could try, if you want to try,

01:17:36,843 --> 01:17:39,741
you can import a few libraries.
You might not need all of them.

This is just a start. I will share the whole thing later.

And because the London one,

if you try to get access to the London bike share data,

it's actually pretty comprehensive.

Let me try to get to it.

Yeah, here.

If you go there,
you should be able to see this bucket loading thing.

Because I was just creating the example,

I got a bit lazy and I saw this ZIP file for 2015 so I just downloaded it and then read it.

It's somewhere. But then you can see that they have data for every single week.

And then, if you scroll down,
you can see the older data in a ZIP file.

So, if you just want to try something to see if it works, just download one of these. And then, what I did is that I set up a folder to house this part.
And then, with this folder,

1306
01:19:16,713 --> 01:19:18,829
I created a file list

1307
01:19:18,888 --> 01:19:24,533
and then I read all these files.

1308
01:19:24,633 --> 01:19:28,347
Because if you actually look at the folder,

1309
01:19:28,615 --> 01:19:34,093
if for the whole 2015 you have one file

1310
01:19:34,193 --> 01:19:35,441
for each week,

1311
01:19:36,190 --> 01:19:39,407
you end up getting a pretty big data set.

1312
01:19:40,295 --> 01:19:42,843
So, yeah, let’s do this.
It might take some time.

After you download it, you can try it.

I'll just put this here in case you want to follow right now.

If you have a fast enough connection.

Of course, this is just my folder path,

which you will have to change.

And then, if you follow this,

then it will be reading these files.
And it will be called London bike share.

If we click into it,

you can see that it's really pretty huge.

I think it's almost 10 million, right?

If I read correctly.

While the one for Nextbike is like 10% of this size, I think,

or even less.

So, it would be more effective
if you can read this using fread,

I suppose, if you're familiar with it.

And then, about the location of the station,

then it gets more interesting

because it actually took me some time to dig it out

and surprisingly, it's actually not from direct download.

It's more like from a freedom of information request,
that people were requesting the location of the stations.

Then I found the file of the locations of stations around that time.

Then I read it.

Then I tried to, because it's a bit messy.

If you can see.

I will just show it very briefly.
You can see that the station name, here you have the station name and then you have some sort of local area in your trip data set. But then, if you look at the station data set, you can see that it's slightly different. So, what I was doing here is just renaming them.
and trying to remove the part of the station name so that they could match.

I tried to extract the part before the comma.

So, afterwards,

you get something that is quite neat and, actually, you could try to arrange it,

you could try to combine them together.

And then, here, this is just something simple,
because if you look back to your station,

to your trip data set,

you can see these names are all separated by spaces

and sometimes spaces could cause problems.

So, I replaced them with dots.

I mean, it's usually up to what you want to do with them.

Or you can replace it with something else.
And then, the next one would be

1368
01:23:12,289 --> 01:23:14,532
just to merge them

1369
01:23:14,889 --> 01:23:17,936
so that we get a different data frame

1370
01:23:21,751 --> 01:23:23,152
that we could use.

1371
01:23:25,727 --> 01:23:29,905
Let's say, I will need the locations of these stations

1372
01:23:30,296 --> 01:23:33,467
to plot them on a map or something,

1373
01:23:34,017 --> 01:23:36,577
that would be pretty useful.

1374
01:23:36,677 --> 01:23:38,899
But the merging would take some time
so I would not run it.

And then, afterwards,

I would like to arrange the number of trips by day.

Actually, this is just to show that

sometimes we might want to do something like

sorting the number of trips by day

or we might want to plot the trip duration by day.

And then, if you think back,
this is actually, if you combine these data sets,

this is actually something that you find a lot

in places like Kaggle

where you can find the number of trips dataset

combined with the weather data set

where people are using these data sets to

run their algorithm from.

So, my point is that
if you get the more complicated data dump,

then you could easily reduce it back

to something that you could use very easily.

Because here, I didn’t include everything.

But then, the next step that I would share would be

something like, after arranging them by day,

I will combine it.
Actually, I find this quite interesting because

1399
01:24:57,600 --> 01:25:02,312
I find daily restriction of

1400
01:25:02,523 --> 01:25:05,641
COVID lockdown restrictions in London.

1401
01:25:06,742 --> 01:25:08,252
So, potentially,

1402
01:25:08,352 --> 01:25:13,754
if I were using the 2020 and 2021 data sets,

1403
01:25:13,881 --> 01:25:17,121
then I could easily run some analysis

1404
01:25:17,632 --> 01:25:22,927
after I combine these for daily number of trips

1405
01:25:23,027 --> 01:25:25,648
and then I also get the station level data,
and also, every day, what the restrictions look like.

This is very interesting.

You can get it from the data for later.

But then, it's also on the Department for Transport site.

The reason that I highlight this is that you can see you have the school closure dummy,

pub closure dummy,

shop closure dummy, and then by date.
And that’s pretty useful information.

If you try to run it and relate it to other data sets, all you need is the date to combine them.

I also mentioned that, in terms of the stations, in terms of what you have between the treatment stations.

Let’s say that you can identify that some stations you have a hypothesis for saying that some stations may be
more affected by these measures than the others.

Then, you can carry out the same type of analysis that I was just talking about earlier on.

Similar to that research paper that I was talking about.

So, this is the idea.

Although, I didn't show every step or the complete codes for now.
But I’m going to share it and I hope that you can see,

01:26:55,214 --> 01:26:57,143
after you get some of this data,

01:26:57,576 --> 01:27:02,336
how you can proceed to get more from them

01:27:02,436 --> 01:27:05,425
by developing a data set

01:27:05,525 --> 01:27:08,263
and a workflow that goes from there.

01:27:08,451 --> 01:27:12,448
And there are numerous research possibilities for that

01:27:12,548 --> 01:27:18,839
as a citizen or even perhaps as a student

01:27:18,939 --> 01:27:22,829
to look through these kinds of things, or even as a researcher, you might
want to explore your research questions with this type of data.

So, this is it for now.

I understand that I didn't really have much time to cover all the steps so you might find it a bit puzzling and a bit chaotic at times.

But I will try to share the materials and the links here so you could try it.
And you could try it later.

So, this is it.

Thanks.