Scripublic transt

Yuan Liao

5/11/2021

Title page

(00:20, 00:20)

Good afternoon, everyone. Thank you for coming to my dissertation defense. In the next 30 mins, I will be talking about my doctoral research: Understanding Mobility and Transport Modal Disparities Using Emerging Data Sources: Modelling Potentials and Limitations.

What is human mobility?

(00:48, 01:08)

So, what is human mobility? It refers to the geographic displacement of human beings, seen as individuals or groups, in space and time.

In Figure 1 on the left side, we can see the individual movements of a few Twitter users over a time period in Stockholm region. Aggregating these movements gives us Figure 2, where we see how central Stockholm is connected with the surrounding municipalities by the trips generated by groups of people.

Understanding human mobility tells us how fast pandemics spread globally, how poverty affects one's travelling behaviour, and where the most attractive places are in a city.

How is human mobility supported by transport systems?

(00:32, 01:40)

Click Mobility happens because people have access to a variety of transport modes, such as ride-sourcing like Uber using our smart phones, public transit, and private car.

Click Different modes have different levels of carbon intensity. They also present distinct characteristics such as travel time and how the trips distribute in space and time. We call these transport modal disparities.

Background

(00:46, 02:26)

Click Transportation presents a major challenge to curbing climate change. Better-informed policymaking requires up-to-date empirical data with good quality, low cost, and easy access. Along with the prevalence of digital technologies, emerging data sources provide us large amount data of human movements and transport systems. How can we use them to gain new insights? This thesis is a practice to answer this question.

Click As examples of emerging data sources, Figure 3 visualises geotagged tweets and road networks of Stockholm region.

Research questions and present work

(01:32, 03:58)

Click This thesis asks two questions. First, what are the potentials and limitations of using emerging data sources for modelling mobility?

Second, how can new data sources be properly modelled for characterising transport modal disparities?

Click The appended papers are organised under two research questions highlighting the use of emerging data sources. However, this thesis thematically tells a single story from understanding to applying. It starts from the population heterogeneity on the fundamental aspects of mobility in Paper I, a more systematic exploration of the data feasibility for travel demand estimation in Paper II, towards a more practical direction — addressing the identified issue of data sparsity with a new model for synthetic travel demand in Paper III.

Then the research continues with putting the movements of people into its context, transport systems by involving more diverse data sources beyond geolocations of people's movements. Paper IV-V provide the insights into the disparities between car and public transit about travel time and modal competition. The results have a high spatiotemporal resolution, and are useful for guiding transport planning and policymaking to make public transit more attractive.

Methodology

(00:33, 04:31)

The methodological framework of this thesis lies in the applied side of data science. Data science is a multi-disciplinary field that intersects between Computer Science/Information Technology, Mathematics and Statistics, and Domain Knowledge which is mobility in this context.

Methods applied by the appended papers are the intersections of each pair of these components. They are data mining, mobility metrics and models, and methods in transport geography.

RQ1 Potentials and limitations of geotagged tweets

(01:00, 05:31) After introducing what this thesis is about, let's take a look at the first research question. What are the potentials and limitations of geotagged tweets for modelling mobility?

[**Click**] Geotagged tweets are the tweets with precise location information when Twitter users actively choose to tag it.

[**Click**] This data source gets increasingly popular due to its easy access, low cost, large spatial and population coverage.

[**Click**] However, it has its limitations. First, Twitter users do not represent the whole population. Based on the research, Twitter users tend to be young, highly-educated, and urban residents.

Moreover, the geolocations of tweets are sparse sampling of the actual mobility and they contain behaviour bias because people selectively report the locations together with their tweets. I will talk about the last two limitations in more detail.

Limitations: sampling of the actual mobility

(01:10, 06:41)

Why do geotagged tweets give incomplete picture of the actual mobility?

[**Click**] Because Twitter users do not geotweet every day and when they geotweet, they do not geotweet every location visited.

[**Click**] These two figures show the statistics of the data collected from users who geotweet most frequently in their countries. Y axis is country and x axes show the share of active days that have at least one geolocation on the left and the number of geotagged tweets per active day on the right.

We see that they have long-tail distributions and differences between countries. Overall, only 20 percent of days have geotagged tweets and for those days, the number of geolocations per active day is only around 1.6. Do you know how many places we visit everyday? In Sweden, it is around 3.1 locations. Now we see how sparse geotagged tweets are in representing the actual mobility.

Limitations: behaviour bias of overly reporting leisure/night activities

(00:40, 07:21)

Besides sparsity, another issue is behaviour bias.

[**Click**] We found that geotagged tweets overly represent leisure or night activities. On the other hand, they under represent regularly visited places such as home and workplace.

[**Click**] These two figures show the share of reported locations over a week for travel survey and geotagged tweets. We can see that Travel survey on the left displays a strong regularity of commuting during weekdays. While geotagged tweets on the right are less regular and increase when the weekend approaches.

Limitations: not for commuting travel demand estimation

(00:59, 08:20)

Given the behaviour bias of reporting uncommon places more than regularly visited places,

[**Click**] it is not surprising to find geotagged tweets are not reliable for commuting travel demand estimation.

[**Click**] On the left side, we have the commuting origin-destination matrices that connect home and workplace at the municipality level in Sweden. For Survey as the ground truth, we see reasonable commuting trips within municipalities or between neighboring ones. However, the matrix using Twitter data has quite a different pattern.

[**Click**] On the right side, we have the commuting trip distance distributions of the two data sources, and Twitter gives the estimation that is far off the one from the travel survey.

Potentials at individual level: Population heterogeneity on mobility

(00:55, 09:15)

Despite those limitations we have discussed so far, population heterogeneity on basic mobility patterns is reflected in geotagged tweets.

[**Click**] We describe mobility with geographical characteristics and network properties. Geographical characteristics reveal how far one travels and network properties tell us how frequently one explores new locations.

Using clustering analysis on these two dimensions, we found four types of travellers: local returner, local explorer, global returner, and global explorer.

[**Click**] Local travellers visit locations that are nearby, while global travellers visit far locations.

[**Click**] Returners tend to visit around one frequently visited centre while explorers travel around decentralised locations.

Now we know geotagged tweets can be used to reveal the individual differences on their travel patterns.

Potentials at population level: travel demand modelling - 1

(00:35, 09:50)

Let's switch to the population level to look at the feasibility of using geotagged tweets for travel demand modelling.

[**Click**] How good Twitter data are for travel demand modelling depends on spatial scale, sampling method, and sample size among others.

[**Click**] Regarding spatial scale, we found Twitter data are more suitable for **city level** than national level. And the main obstacle of using Twitter data at a large spatial scale is the **sparsity**.

Potentials at population level: travel demand modelling - 2

(01:01, 10:52)

[**Click**] Regarding sampling method, we found **user-based** data collection works better than area-based data collection, because it gives a much larger number of geotagged tweets and therefore, a more complete picture of travel demand.

[**Click**] To make more data usable, we propose a density-based approach.

[**Click**] In the literature, two consecutive geotagged tweets by one person make a trip as shown in the left side. However, we already know that Twitter users do not geotweet everyday or every location. So, this trip-based way needs to filter out a lot of trips that have unreasonably long time interval. This makes the sparse Twitter data even more sparse. In this density-based approach on the right side, we use population and the count of geotagged tweets to decide the trips between traffic zones. It increases sample size a lot and results in a more robust and accurate estimation of the overall travel demand.

Extending the use by innovative approaches - 1

(01:08, 12:00)

The more complete the data, the better picture of human mobility they provide. Given the sparsity of geotagged tweets, the feasibility of using them is limited. However, we can extend the use by innovative approaches.

[**Click**] We have talked about the density-based approach for travel demand estimation which gives better results by using more data. In paper III, we further develope an individual-based mobility model that fills the gaps in geotagged tweets. The model is designed to correct behaviour bias and sparsity issue we have discussed.

[**Click**] The model takes the input of sparse mobility traces that can not be directly converted to trips, because they are sparse in time. As we can see on the left side. The model uses the fundamental mechanisms of mobility to synthesise mobility traces based on individual data, so the output can be converted to representative daily trips, as shown on the right side.

Extending the use by innovative approaches - 2

(00:40, 12:40)

[**Click**] The model-synthesised results have good agreements with the other data sources such as travel survey.

[**Click**] We demonstrate an application here. We can use the model to characterise trip distance distributions of global regions' residents. Here we see the model-synthesised results in Figure 11. X axis is trip distance and Y axis is probability density. From left to right, we have Sweden, Nairobi, and all the regions. We see some regional differences but they mostly follow a lognormal distribution. We can use these synthesised data for travel demand estimation which can scale up due to the easy access of social media data.

RQ2 Transport modal disparities

(01:09, 13:49)

After showing you some results in answering research question 1, let's put human movements in their context, transport systems. The second research question of this thesis looks into transport modal disparities.

As shown in this diagram, Paper IV is about spatial temporal disparities of travel time between car and public transit. Paper V is about how the built environment and trip attributes affect the competition between ride-sourcing and public transit.

Regarding the involved data sources, Paper IV uses more than geolocations of Twitter; we also integrated road traffic, networks, and public transit schedules from transport systems. Paper V further extends the data use by combining ride-sourcing trip data, points of interest, and weather records.

In the upcoming slides, I will present some results in quantifying transport modal disparities between car and public transit as well as ride-sourcing and public transit.

Spatiotemporal patterns of travel time - 1

(00:37, 14:26)

[Click] In Paper IV, we propose a data fusion framework of travel demand, network operations, and infrastructure. The data fusion framework allows us to combine both demand and operations while getting the results of high resolution, especially through the use of Twitter data as a proxy for time-varying travel demand.

[Click] As shown in the animation, the distribution of geotagged tweets represents the attractiveness of various locations in cities and they reveal the dynamics of visits over the course of a day.

Spatiotemporal patterns of travel time - 2

(00:51, 15:17)

[Click] We define the travel time ratio as the travel time by public transit divided by the travel time by car. A high modal disparity means taking public transport is much more time-consuming than driving a car. Here we see the 24-hour dynamics of the travel time disparity. The results talk about when and where to improve public transt service in terms of both access and travel time.

[Click] We further summarise the travel time disparity across cities which varies over 24 hours, and find that travel time by public transt is around twice as high as by car. It turns out that public transt can compete with car use during peak rush hours in Stockholm and Amsterdam.

Modal competition: ride-sourcing vs. public transit - 1

(01:24, 16:41)

People can choose driving their own car or taking public transit. Nowadays, ride-sourcing like Uber has become a popular alternative to driving.

[Click] One of the key questions remains unanswered: Does ride-sourcing complement, or compete with, public transit? Public transit has lower green house gas intensity than ride-sourcing. Therefore, if ride-sourcing mostly competes with public transt, it may increase the overall emissions from transport systems.

[Click] Let's ask ourselves this question: How large is the share of ride-sourcing trips that can be substituted by taking public transit? If you are willing to walk up to 800 m to access and leave the transit station during daytime?

[Click] By this definition, we call the ride-sourcing trips that can be substituted by public transit transit-competing, while the ones that can not be substituted as non-transit-competing. What trip attributes and built environment are linked to the competition? What are the implications for policymaking? Before we approach these questions, let's look at the data at our hands.

Modal competition: ride-sourcing vs. public transit - 2

[Click] It turns out that the transit-competing trips account for 48.2% of the total trip records. It means that a considerable share of ride-sourcing trips can be potentially substituted by public transit.

[Click] Looking at their hot spots of pick-up and drop-off locations, the non-transit-competing trips tend to have a more spread-out distribution of pick-up and drop-off hot spots, including the international airport.

Data fusion approaches

(02:12, 18:53)

[Click] We have demonstrated a data fusion framework for travel time calculation. For a better understanding of transport modal disparities, we need innovative ways of utilising different data sources, especially increasing the amount of incomplete but big datasets that are made publicly available. These data can cover trips from ride-sourcing, ride-sharing, taxi, and e-scooters.

[Click] They are oftentimes collected from a large area and population but at a cost of rich detail. A common set of variables in these big trip data include trip ID, pick-up and drop-off locations, pick-up and drop-off times, and cost. To gain insights from using these data, we need a data fusion framework to enrich the original dataset.

[Click] In Paper V, we enrich each record with the travel information for its public transt alternative assuming the same departure time, origin, and destination. We also get the weather information for the departure time. Moreover, we detect the community structure of the ride-sourcing origin-destination matrix. By doing so, we divide the study area into sub-regions based on the ride-sourcing travel demand. These demand-based sub-regions help us better identify where the competition between them.

The original dataset does not tell us any environmental context. In order to know more about the built environment of the pick-up and drop-off spots, we identify the functional clusters using points of interest (POIs) of the zones in the study area, such as restaurant and hotel, and quantify the transit-stop density in the zones.

After the data enrichment, the ride-sourcing trips are ready for the analysis. Now we can answer how the competition is affected by the trip attributes and built environment.

Modal competition: ride-sourcing vs. public transit

Model and impact of travel time and public-transit transfer

(01:15, 20:08)

[Click] **[Click]** Besides the data fusion framework, we also use a glass-box model enhanced by machine learning techniques with the enriched ride-sourcing trips to discuss the relationship between the two modes. We quantify how the many factors of trip attributes and built environment and the interactions between them affect the competition between them in an additive way.

[Click] This figure shows the impact of travel time by ride-sourcing and transfer of taking public transit. The y axis shows their scores; a score above zero increases the tendency of a trip being transit-competing.

[Click] We found that Competition is more likely to happen when the **travel time** by ride-sourcing < 15 min. Requiring multiple **transfers** is associated with the competition between.

Modal competition: ride-sourcing vs. public transit

Impact of land-use patterns

(00:54, 21:02)

[Click] **[Click]** Using points of interest such as hotel and restaurant, we group the study area into seven clusters. They are named by their land-use intensity and diversity. From rural to centre, the land-use diversity and intensity increase.

[Click] This figure shows the impact of the land-use cluster on the left and its interaction with public transit transfer on the right. It turns out that low density/diversity of land use lowers the probability of competition between the two modes. And interestingly, multiple transfers required by the trips in the area of middle density/diversity land use intensify the competition.

Modal competition: ride-sourcing vs. public transit

Selective recommendations

(00:43, 21:45)

[Click] Based on the main findings, I present some recommendations here.

[Click] We could improve public transt services that provide access to the international airport.

[Click] Expand public transt networks guided by the transit-competing ride-sourcing trips featuring short travel time but a big gap between the two modes.

[Click] We could also incentivise the ride-sourcing trips that fill the gaps in the public transt services that take a long time or require lengthy walking and transfers connecting to suburban areas.

A summary of answers

RQ1 Potentials and limitations for modelling mobility

(00:53, 22:38)

[Click] **[Click]** After selectively showing you some results from the five studies, let's summarise the answers to research question one.

[Click] Regarding the use of social media data, we need to balance the trade-offs between easy access, low cost, and biased population, behaviour bias, and sparsity issue.

[Click] At the individual level, fundamental patterns are preserved such as mobility regularity, diffusive nature, and returning effect as well as population heterogeneity.

[Click] At the population level, geolocations of Twitter data is reasonably good for the overall travel demand estimation but not commuting demand. However, we need carefully consider spatial scale, sampling method, and sample size.

[Click] At last, we can extend the use by innovative approaches to correct the biases and increase the available data.

A summary of answers

RQ2 Characterising transport modal disparities

(00:55, 23:33)

[Click] **[Click]** Putting human movements into transport systems, what do we find about the disparities between public transit and car and ride-sourcing?

[Click] We demonstrate the importance of data fusion approaches, especially given more and more open but incomplete data are available. For instance, in estimating travel time disparities, we introduce geotagged tweets.

[Click] Which turns out a good source for time-varying attractiveness of urban locations.

[Click] It is not surprising to find that public transit is virtually always slower than car and ride-sourcing.

[Click] However, for making public transit more competitive, these spatiotemporal details add nuanced insights to identify gaps and opportunities for policymaking and transport planning.

Outlook - 1

(01:21, 24:54)

Paper I to V use emerging data sources in mobility and transport from the exploratory stage to the application side. After four-year research, compared with much more knowledge still needed, this doctoral research is just a start. I believe the future work should ask relevant questions with powerful and innovative tools without constraining the research into specific data sources. There are four directions I'd like to pursue in the future.

[Click] The first direction is to extend the use of social media data for mobility modelling. Besides the continuous effort of debiasing this data source, long-distance travels could be a reasonable research object using this data source. Becayse we can get many leisure activities and international travels from it.

This thesis focuses on the geolocations of Twitter data, while the text part is rich in information and can provide better context together with the geolocation part.

[Click] Towards the end of this doctoral research, the studies have become more cross-disciplinary and practical. What we just started is to generate global synthetic mobility data for improving travel demand projections. This is to be integrated into the energy systems' modelling for the transport sector.

Outlook - 2

(01:18, 26:12)

The transport sector accounts for a big share of carbon emissions. It is valuable to seek concrete policy implications of minimising the carbon footprint from the transport sector in cities.

[Click] One way of doing this is to combine multi-modal trip data for occupancy, shareability, and electrification of new mobility services provided by TNCs. **[Click]** This thesis has used a series of metrics and models from network science. However, the perspective of networks has only been implicit, and the use of network science tools has been superficial and practical. I found the perspective of networks interesting and powerful to reveal the patterns of urban mobility that no other tools are capable of.

One direction that I haven't had time to pursue is to study the relationship between user/traveller friendship networks (abstract) and their mobility networks (spatial). This tells us about social segregation and how such segregation and spatial interactions shape each other. Along this direction, advanced techniques from graph learning and complex systems will contribute to urban mobility research together with the increasingly available data.

Thanks for listening!

(00:20, 26:32)

Thanks for listening. That's all for the presentation part. You can reach me via email, Twitter, or my personal website. You can access the thesis using this link or using your phone to scan the QR code to visit it.