Script

Yuan Liao

06/03/2021

Title page

(21)

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.

Background

(43)

[**Click**] Transportation presents a major challenge to curbing climate change.

[**Click**] Better-informed policymaking requires up-to-date empirical data with good quality, low cost, and easy access.

[**Click**] 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**] There two key concepts here, mobility and transport modal disparities which I will explain in detail.

What is human mobility?

(48)

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

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

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?

(36)

[Click] Mobility is supported by 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.

Research questions and present work

(110)

[Click] This thesis asks two questions. First, what are the potentials and limitations of using emerging data sources, particularly geotagged social media data, 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

(38)

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 What are the potentials and limitations of using geotagged tweets for modelling mobility?

(27)

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?

In this part, I will introduce the main findings from the first three appended papers for both limitations and potentials of using this particular data source.

RQ1 Geotagged tweets, pros and cons

(54)

[**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.

We also confirm that the geolocations of tweets are sparse sampling of the actual mobility and they contain behaviour bias because Twitter users selectively report the locations. I will talk about how we reveal the last two limitations in our studies.

RQ1 Limitations: sparse sampling of the actual mobility | (Paper III)

(73)

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

[**Click**] Because Twitter users do not geotweet every day.

Based on the data collected from users who geotweet most frequently in their countries, only 20 percent of days have geotagged tweets.

[**Click**] When Twitter users geotweet, they do not geotweet every location visited.

For those active 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.

RQ1 Limitations: behaviour bias of overly reporting leisure/night activities | (Paper I)

(61)

Besides sparsity, another issue is behaviour bias.

[**Click**] We confirm 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 visited 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 reported locations increase when the weekend approaches.

RQ1 Limitations: not for commuting travel demand estimation | (Paper II)

(45)

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, which is one of the main findings from Paper II.

[Click] 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 as ground truth.

RQ1 Potentials at individual level: population heterogeneity on mobility | (Paper I)

(66)

I have shown you those limitations we have revealed in our studies. For the next few slides, I will talk about the revealed potentials of using this data source.

Let's look at the population heterogeneity on basic mobility patterns.

[Click] In Paper I, 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 reasonable individual differences on their travel patterns.

RQ1 Potentials at population level: travel demand modelling | (Paper II)

(55)

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

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.

[Click] Regarding sampling method, we found **user-based** data collection works better than area-based data collection,

[Click] because it gives a much larger number of geotagged tweets and therefore, a more complete picture of travel demand.

RQ1 Extending the use by innovative approaches | (Paper II)

(60)

[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 unreasonable trips. This makes the sparse Twitter data even more sparse.

[Click] 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.

[Click] It increases sample size and gives more robust and accurate estimation of the overall travel demand.

RQ1 Extending the use by innovative approaches | (Paper III)

(90)

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. Therefore, we propose approaches to extend the use.

We have talked about the density-based approach for travel demand estimation which gives better results by using more data.

[Click] 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] As we can see on the left side, the model takes the input of sparse mobility traces that can not be directly converted to trips, because they are sparse in time.

[Click] As shown on the right side, the model uses the fundamental mechanisms of returning and exploring to synthesise mobility traces, so the output can be converted to representative daily trips.

RQ1 Extending the use by innovative approaches | (Paper III)

(45)

[Click] The model-synthesised results have good performance.

[Click] I'd like to show you an application here: characterise trip distance distributions of global regions' residents. Here we see the model-synthesised domestic trips for a list of regions.

X axis is trip distance and Y axis is probability density. We can use these synthesised data for travel demand estimation. This provides us a solution which can scale up, due to the easy access of social media data.

A summary of answers to RQ1 | Potentials and limitations of geotagged tweets for modelling mobility

(70)

[Click] **[Click]** After selectively showing you some results from answering RQ1, let's recap the answers.

[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.

RQ2 How can new data sources be properly modelled for characterising transport modal disparities?

(22)

Now, let's put human movements in their context, transport systems. The second research question looks into how new data sources can be properly modelled for characterising transport modal disparities?

In this part, I will talk about the main findings from the last two appended papers.

RQ2 Spatiotemporal patterns of travel time: data fusion approach | (Paper IV)

(49)

[Click] In Paper IV, the main contribution is that we propose a data fusion framework combining travel demand, network operations, and infrastructure. The data fusion framework allows us to calculate travel time with high spatiotemporal 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.

RQ2 Spatiotemporal patterns of travel time | (Paper IV)

(65)

[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.

RQ2 Modal competition: ride-sourcing vs. public transit | (Paper V)

(113)

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.

RQ2 Modal competition: ride-sourcing vs. public transit | (Paper V)

(30)

[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.

RQ2 Modal competition: data fusion approach and model | (Paper V)

(136)

[Click] Nowadays, increasingly amount of incomplete but big datasets that are made publicly available, such as ride-sourcing trip data.

They are oftentimes collected from a large area and population but at a cost of rich detail.

[Click] In paper V, the raw data only 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 to enrich the original dataset.

[Click] We enrich each record with public transit travel information, weather information, and demand-based communities by using the spatial network of all the ride-sourcing trips.

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 add the transit-stop density in the zones. We also identify the functional clusters using points of interest (POIs) of the zones in the study area, such as restaurant and hotel.

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.

[Click] For doing so, we use a glass-box model enhanced by machine learning techniques to quantify the impact of the variables and the variable interactions on the competition between the two modes.

RQ2 Modal competition: impact of land-use | (Paper V)

(85)

[Click] Using points of interest such as hotel and restaurant, we group the study area into seven clusters. From rural to centre, as color from blue to red, the land-use diversity and intensity increase.

[Click] This figure shows the impact of the land-use functional cluster. A score above zero means that having a drop-off location of a certain functional cluster increases the probability of the competition, and vice versa.

[Click] It turns out that low density/diversity of land use lowers the probability of competition between the two modes.

RQ2 Modal competition: impact of land-use x transit boardings | (Paper V)

(60)

[**Click**] [**Click**] This figure shows the impact of the interaction between land-use cluster and public transit boardings.

[**Click**] We found that multiple transfers required by the trips in the area of middle density/diversity land use intensify the competition.

RQ2 Modal competition: selective recommendations | (Paper V)

(48)

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

[**Click**] We could expand public transt networks guided by the transit-competing ride-sourcing trips featuring short travel time but a big travel time disparity 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 or require transfers connecting to suburban areas.

A summary of answers to RQ2 | Characterising transport modal disparities between public transit and car & ride-sourcing

(80)

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.

Knowledge contributions

(54)

[**Click**] [**Click**] This thesis provides validation to identify the limitations of geotagged tweets: **behavior bias**, **biased population**, and **sparsity issue**. Especially we highlight the unreliability of using it for commuting demand.

[**Click**] It reveal the potentials of geotagged tweets at both **individual** and **population** level.

[**Click**] It also reveals the spatio-temporal disparities between car/ride-sourcing and public transit about **travel time** and **modal competition**.

The knowledge is useful for transport planning and future mobility research of continuously using geotagged tweets.

Methodological contributions

(40)

[Click] **[Click]** This thesis proposes **a density approach** and **an individual-based mobility model** for travel demand estimation, addressing sparsity issue and behaviour bias of geotagged tweets.

[Click] It also creates two reproducible **data fusion frameworks** integrating multiple data sources from transport systems for characterising modal disparities.

Future mobility research can apply the frameworks to different regions with different datasets.

Outlook

(110)

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.

[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.

[Click] The transport sector accounts for a big share of carbon emissions. It is valuable to seek concrete policy implications of minimising its carbon footprint.

One way of doing this is to combine multi-modal trip data to explore the issues of 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. The perspective of networks interesting and powerful to reveal the patterns of urban mobility that no other tools are capable of.

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!

(22)

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.