THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Understanding Mobility and Transport Modal Disparities Using Emerging Data Sources: Modelling Potentials and Limitations

YUAN LIAO

Department of Space, Earth and Environment
Division of Physical Resource Theory
CHALMERS UNIVERSITY OF TECHNOLOGY

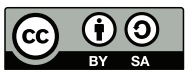
Göteborg, Sweden 2021

Understanding Mobility and Transport Modal Disparities Using Emerging Data Sources: Modelling Potentials and Limitations

YUAN LIAO
ISBN 978-91-7905-508-0

ISSN 0346-718X

Department of Space, Earth and Environment
Division of Physical Resource Theory
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone: +46 (0)31-772 1000



Chalmers Reproservice
Göteborg, Sweden 2021

ABSTRACT

Transportation presents a major challenge to curb climate change due in part to its ever-increasing travel demand. Better informed policy-making requires up-to-date empirical mobility data to model viable mitigation options for reducing emissions from the transport sector. On the one hand, the prevalence of digital technologies enables a large-scale collection of human mobility traces, providing big potentials for improving the understanding of mobility patterns and transport modal disparities. On the other hand, the advancement in data science has allowed us to continue pushing the boundary of the potentials and limitations, for new uses of big data in transport.

This thesis uses emerging data sources, including Twitter data, traffic data, OpenStreetMap (OSM), and trip data from new transport modes, to enhance the understanding of mobility and transport modal disparities, e.g., how car and public transit support mobility differently. Specifically, this thesis aims to answer two research questions: (1) What are the potentials and limitations of using these emerging data sources for modelling mobility? (2) How can these new data sources be properly modelled for characterising transport modal disparities? Papers I-III model mobility mainly using geotagged social media data, and reveal the potentials and limitations of this data source by validating against established sources (Q1). Papers IV-V combine multiple data sources to characterise transport modal disparities (Q2) which further demonstrate the modelling potentials of the emerging data sources (Q1).

Despite a biased population representation and low and irregular sampling of the actual mobility, the geolocations of Twitter data can be used in models to produce good agreements with the other data sources on the fundamental characteristics of individual and population mobility. However, its feasibility for estimating travel demand depends on spatial scale, sparsity, sampling method, and sample size. To extend the use of social media data, this thesis develops two novel approaches to address the sparsity issue: (1) An individual-based mobility model that fills the gaps in the sparse mobility traces for synthetic travel demand; (2) A population-based model that uses Twitter geolocations as attractions instead of trips for estimating the flows of people between regions. This thesis also presents two reproducible data fusion frameworks for characterising transport modal disparities. They demonstrate the power of combining different data sources to gain new insights into the spatiotemporal patterns of travel time disparities between car and public transit, and the competition between ride-sourcing and public transport.

Keywords: mobility models, social media data, traffic data, big trip data, transport modes, data mining, geographical information systems

APPENDED PUBLICATIONS

This thesis consists of an extended summary and the following appended papers:

Paper I Y. Liao, S. Yeh and G. S. Jeuken (14th Nov. 2019). From individual to collective behaviours: exploring population heterogeneity of human mobility based on social media data. *EPJ Data Science* **8** (1), p. 34. DOI: 10.1140/epjds/s13688-019-0212-x.

Paper II Y. Liao, S. Yeh and J. Gil (26th Jan. 2021). Feasibility of estimating travel demand using geolocations of social media data. *Transportation*, pp. 1–25. DOI: 10.1007/s11116-021-10171-x.

Paper III Y. Liao, K. Ek, E. Wennerberg, S. Yeh and J. Gil (26th Apr. 2021). *A mobility model for synthetic travel demand from sparse individual traces.* Submitted to Computers Environment and Urban Systems, Under review.

Paper IV Y. Liao, J. Gil, R. H. M. Pereira, S. Yeh and V. Verendel (4th Mar. 2020). Disparities in travel times between car and transit: spatiotemporal patterns in cities. *Scientific Reports* **10** (4056). DOI: 10.1038/s41598-020-61077-0.

Paper V Y. Liao (11th May 2021). *Ride-sourcing compared to its public-transit alternative using big trip data.* Submitted to Journal of Transport Geography, Under review.

Author contributions

Paper I: YL, SY designed the study. YL analysed the data. YL and SY wrote the paper. All authors edited and approved the final version of this manuscript.

Paper II: YL and SY conceptualised the study. YL, JG, and SY designed the methods. YL analysed the data. All authors wrote the manuscript.

Paper III: YL conceptualised the study. YL, JE, and EW designed the methods. YL, JE, and EW analysed the data. All authors wrote the manuscript.

Paper IV: YL, JG, and SY conceptualised the study. YL and VV preprocessed the data. YL, JG, and RP processed and analysed data. All authors wrote the manuscript.

Paper V: YL designed the study, analysed the data, and wrote the manuscript.

OTHER PUBLICATIONS

Other publications by the author not included in the thesis:

Y. Liao and S. Yeh (10th Dec. 2018). 'Predictability in human mobility based on geographical-boundary-free and long-time social media data'. In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 2068–2073. DOI: 10.1109/ITSC.2018.8569770.

Y. Liao, M. Wang, L. Duan and F. Chen (26th Apr. 2018a). Cross-regional driver–vehicle interaction design: an interview study on driving risk perceptions, decisions, and adas function preferences. *IET Intelligent Transport Systems* **12** (8), pp. 801–808. DOI: 10.1049/iet-its.2017.0241.

Y. Liao, G. Li, S. E. Li, B. Cheng and P. Green (28th June 2018b). Understanding driver response patterns to mental workload increase in typical driving scenarios. *IEEE Access* **6**, pp. 35890–35900. DOI: 10.1109/ACCESS.2018.2851309.

M. Wang, Y. Liao, S. L. Lyckvi and F. Chen (2020). How drivers respond to visual vs. auditory information in advisory traffic information systems. *Behaviour & Information Technology* **39** (12), pp. 1308–1319. DOI: 10.1080/0144929X.2019.1667439.

G. Li, Y. Liao, Q. Guo, C. Shen and W. Lai (28th Jan. 2021). Traffic crash characteristics in shenzhen, china from 2014 to 2016. *International Journal of Environmental Research and Public Health* **18** (3), p. 1176. DOI: 10.3390/ijerph18031176.

G. Li, S. E. Li, R. Zou, Y. Liao and B. Cheng (Oct. 2019). Detection of road traffic participants using cost-effective arrayed ultrasonic sensors in low-speed traffic situations. *Mechanical Systems and Signal Processing* **132**, pp. 535–545. DOI: 10.1016/j.ymssp.2019.07.009.

To my family
&
to my love – Jian – and the new adventure we are heading off together

ACKNOWLEDGEMENTS

Four years is short. I still remember how I started my first day at the division; it is like yesterday to me. My doctoral research has been an adventure since I decided to switch the research direction from vehicular human factors to mobility data science. This thesis would not have been possible without the support from my supervisors, co-authors, colleagues, and families.

Thank you, my supervisor and co-author Sonia Yeh for your trust, patience, and inspiring guidance. You always ask sharp and hard questions that drive me forward and make me see a different world. You are an excellent, open-minded mentor, reliable research partner, and my life-long role model. Thanks to Jorge Gil, my co-supervisor, and co-author. You joined the journey after one and a half years, though I feel you are always there to support me. I admire and learn from your passion for research. The discussion with you is always helpful and critical.

Thank you, my examiner, Kristian Lindgren who contributed great thoughts on my research, and my manager Martin Persson who taught me how to better balance research and life. Thanks to Daniel Johansson, I appreciate your support on both my work and research. Great thanks to my co-authors, Rafael H. M. Pereira, Gustavo S. Jeuken, Vilhelm Verendel, Kristoffer Ek, and Eric Wennerberg. It's been a pleasure to work with you and I learned a lot from you. Kristoffer and Eric, you are the first two master students that I worked with at Chalmers. Thanks for trusting me and the experience of coding in the same project has permanently increased my productivity.

Big thanks to all friends and colleagues in the division of Physical Resource Theory for providing a fantastic working environment and supports during hard times like the COVID-19 pandemic. Special thanks to my office mates, Xiaoming Kan, Ella Rebalski, and Ahmet Mandev who make my work full of joy. I miss the old times when we could still meet, chat, and laugh every day.

Thanks to my family for supporting and encouraging me, through successes and failures along this journey, though you wonder when I will stop being a student and start a real job. Special thanks to Jian, my love, my gold teammate, my everything. We are together for eleven years now. I always imagine I would watch you get a PhD sometime in the future, because you are so smart, patient, and always endeavouring for the perfect. When I stand in this position, I just want to say that without your support, I would not have made it this far.

At last, thanks to my unborn baby for the kicks which keep inspiring me in writing this thesis. I wish you healthy and happy. We can't wait to hold your hands to explore this world together.

Göteborg, May 2021
Yuan Liao

CONTENTS

CHAPTER 1

Introduction

Human mobility refers to the geographic displacement of human beings, seen as individuals or groups, in space and time. The study of this important subject spans several disciplines, e.g., complex systems [1] and transport geography [2]. The study outcomes have a broad relevance; they reveal how fast epidemics spread globally in epidemiology [3], they show how poverty affects one's travelling behaviour in social science [4], and they tell us where the most attractive places are in a city in transport planning [5].

The transport sector, i.e., the movement of people and goods, is likely to become the sector with the highest emissions in the 2°C scenario after 2030 [6]. There are many ways to reduce the carbon emissions in the transport sector, where people movements account for a big portion (54%)[1]. For example, policymakers worldwide recognise the importance of promoting a mode shift from car to public transit and other low-carbon modes in cities. Better informed and timely policymaking requires up-to-date empirical data with good quality to create a better understanding of disparities between different transport modes. However, the conventional methods of mobility data collection such as household travel survey have increased cost while the response rates are becoming lower over time [7].

Along with the sea-change development of Information and Communication Technologies (ICT), a large-scale collection of human mobility traces and urban-sensed data in the transport sector has become feasible.

Mobility traces can be collected through online social media platforms e.g., Twitter, GPS-enabled devices, smart card, and call detail records (CDR). Unlike the data collected through household surveys, these emerging data sources of mobility traces have unique features, including the passive collection, large volume, easier access, incompleteness such as no trip purpose and social demographic information, and potential selective bias. Despite some disadvantages, these emerging data sources contribute significantly to both the understanding of mobility using physical models and applications in the field of transport. For example, to what extent human mobility is predictable has been quantified using GPS trajectories [8]. However, their

[1] Source: International Energy Agency (IEA) and the International Council on Clean Transportation (ICCT), 2018

potentials and limitations for further application are not fully explored.

Among these emerging data sources of mobility traces, social media data become especially appealing due to its low cost and easy access. The main criticism against using this type of data pertains to two aspects, a biased population representation and low and irregular sampling of the actual mobility trajectories. There is a consensus on the need for careful inspection when using geotagged social media data to approximate the actual travel behaviours for the general population [9]. Therefore, besides the attempts to gain new insights into human mobility using social media data, modelling its potentials and limitations via the validation against the other data sources is one of the key aspects explored in this thesis.

The other stream of emerging data sources, Geographical Information Systems (GIS) data, especially open GIS data, are the measurement of transport systems where mobility happens. Common data are open road network data, e.g., OpenStreetMap (OSM), public transit schedules and routes, e.g., General Transit Feed Specification (GTFS) data, traffic speed data, e.g., HERE traffic, and mode-specific GPS data, e.g., taxi trip data. A better understanding of transport modal disparities calls for innovative ways of utilising these different data sources, especially increasing amount of incomplete but big datasets are made publicly available such as mode-specific trip data. Along this direction, this thesis also looks into the development of reproducible data fusion frameworks that combine mobility data and GIS data to better understand the disparities between public transit and car driving regarding spatio-temporal patterns of travel time and modal competition.

Scope and contributions

This thesis uses a variety of emerging data sources, including Twitter data, traffic data, and mode-specific trip data, to enhance the understanding of mobility and transport modal disparities. Specifically, this thesis aims to answer two research questions:

- What are the potentials and limitations of using these emerging data sources for modelling mobility?
- How can these new data sources be properly modelled for characterising transport modal disparities?

In answering the above questions, Papers I-III model mobility mainly using geotagged social media data revealing their potentials and limitations (Q1). Papers IV-V combine multiple data sources to characterise transport modal disparities (Q2) which further demonstrate the modelling potentials of the emerging data sources (Q1). Using geotagged Twitter data, Paper I [10] reveals the population heterogeneity of mobility patterns. Paper II [11] examines the effects of data sparsity, spatial scale, sampling methods, and

sample size on the feasibility of using geolocations of social media data for travel demand estimation. Sparse individual mobility traces collected from call detail records and social media platforms have been widely used to study mobility patterns while the sparsity issue has been generally ignored. Paper III [12] extends the use of these inexpensive and easy-to-access data, by proposing a model to fill the gaps in sparse individual traces for synthetic travel demand. In Paper IV [13], we combine multiple sources of information: the travel demand as revealed by Twitter data, the transportation network, and historical road speed records, at a high spatial and temporal granularity. From such a data fusion framework, we provide a more realistic picture of the modal disparity in travel time between car and public transit in four cities in different countries. Paper V [13] uses a public ride-sourcing trip dataset to explore the potential competition between ride-sourcing and public transit as its alternative, and what trip attributes and built environment types are associated with the competition.

The first aspect examined by the thesis is the potentials and limitations for modelling mobility using Twitter data. When validating against the other mobility data sources, Twitter data are representative when the individuals represent the overall population and the key mobility indicators show a small discrepancy, e.g., trip distance, travel demand (represented by the origin-destination matrix), and temporal profiles of activities. Despite having clear signs of overly representing residents living in big cities and their leisure activities, geotagged tweets preserve mobility regularity, diffusive nature, and preferential return to some extent. Paper I illustrates that the fundamental patterns of population heterogeneity on mobility are well preserved in Twitter data. In addition, Paper II sheds light upon a more practical direction: geotagged tweets contribute to a reasonably good travel demand estimation with stability over time. However, the potential for estimating travel demand using geolocations of Twitter data is affected by spatial scale, data sparsity, sampling method, and sample size. We can extend the use of such a low-cost and easy-to-access data source by overcoming the sparsity issue. To do so, this thesis develops two innovative approaches in Papers II and III.

The second aspect examined by the thesis is characterising transport modal disparities using emerging data sources. This thesis features the use of data fusion approaches in both Paper IV and V that contribute to innovative ways of utilising different GIS data sources. Informed by such data fusion approaches, the high spatiotemporal depiction of travel time shows that using PT takes on average 1.4 – 2.6 times longer than driving a car (Paper IV). The share of the area where travel time favours PT over car is surprisingly small at a magnitude of 1% consistently across four cities. Regarding the competition between ride-sourcing (by car) and PT, there is no doubt that a large share of ride-sourcing could have been done by PT (Paper V). Given the significant emission benefits of taking PT instead of

ride-sourcing or car in general, making PT competitive becomes important. To this end, this thesis sheds light upon how PT could be improved to be more attractive to travellers (Paper V). For instance, one could decrease the travel time by PT, especially for those ride-sourcing trips for which TT by ride-sourcing is less than 15 min, and decrease the number of transfers by increasing the connectivity between frequently connected pairs of zones, especially for the trips connecting outer rings of the cities. However, it remains to further investigate whether the implications are applicable to the other regions outside the study area.

The methodological contributions of this thesis lie in the applied side of data science in physics and transport, including data mining, mobility metrics and models, and methods in transport geography. The application of clustering provides new insights into the population heterogeneity of mobility (Paper I). A glass-box model for classification, enhanced by machine learning techniques, is used to discuss the relationship between ride-sourcing and its public-transit alternative, producing intelligible results (Paper V). This thesis demonstrates two reproducible data fusion frameworks for combining different GIS data sources (Papers IV and V). To address sparsity issue of social media data, two new models are proposed: a density-based approach to produce origin-destination matrices (Paper II), and an individual-based mobility model that fills the gaps in the sparse mobility traces (Paper III).

Disposition of this thesis

The thesis consists of five chapters providing brief introduction to my doctoral research, followed by five appended papers. Chapter 2 gives further background on human mobility: how is it defined, how is it facilitated by the transportation systems, what are the emerging data sources that deepen our understanding of it and its interaction with transport modes? Chapter 3 positions my doctoral research in data science, provides an overview of the methodological framework, and it gives a brief literature review of the relevant methods with the focus on the ones applied in the appended papers. Chapter 4 and Chapter 5 summarise and discuss the five appended research papers. Chapter 6 synthesises the main findings, and gives reflections on my doctoral research and an outlook into the future directions of further using emerging data sources in mobility and transport.

CHAPTER 2

Background

This chapter first gives an overview of human mobility on its definition, scope, and applications (Section 2.1). Human mobility is supported by transport systems. Section 2.2 reviews the diversity of modes provided by the transport systems, particularly public transit and car. To better understand mobility and transport modal disparities, empirical data have been widely applied. In the last section of this chapter (Section 2.3), we introduce some emerging data sources measuring mobility traces and transport systems, in comparison with conventional data sources.

2.1 Defining human mobility

Human mobility refers to the geographic displacement of human beings, seen as individuals or groups, in space and time. This displacement constitutes of an origin, a destination, and a specific trajectory in between. Here I give three ways of categorising mobility originated from different disciplines.

Social scientists categorise this mobility (spatial mobility) by its utility [14]: (1) mobility that happens inside the place of residence; (2) migration (international and inter-regional mobility); (3) travel with the purpose of tourism or business; and (4) day-to-day journeys such as commuting and running errands.

Physicists describe mobility by spatiotemporal scale: long-term mobility that is likely to cover large displacement, e.g., migration, and short-term mobility whose displacement is constrained by 24 hours in a day, e.g., commuting. They see mobility as a diffusion process that is characterised by both randomness and regularity [1].

In transport geography, researchers see the mobility as individual behaviour that formulates flows of population. At the individual level, the mobility trajectory is a time series of visits to various locations. Individuals' mobility trajectories can be aggregated to study the flows of people travelling between different locations/regions. Depending on the spatiotemporal scale of the aggregation, an origin-destination matrix (ODM) can be constructed with the origins and the destinations of all trips. An ODM quantifies the population travel demand in a certain area.

In the study of human mobility, quantitative theory seeks to answer relevant questions [15]. Why does an individual start a trip at a certain time? What are the factors that decide the mode choice of travellers? Which route does one choose and why? To what extent is the mobility predictable? The answers to these questions provide insights for a wide range of disciplines, including urban planning [16], transport management [17], epidemiology [18], ecology, and social science [19].

2.2 Moving in the transport systems: public transit vs. car

To study how mobility is supported by transport systems, we need to first understand what transportation is about. According to the definition in Collins Dictionary, "transportation is a system for taking people or goods from one place to another, for example using buses or trains." Regarding transportation as being studied, William R. Black states:

> "Transportation is concerned with the movement of goods and people between different locations and systems used for this movement. Included in the former would be the journey to work, trade flows between nations, commodity flows within a single nation, passenger flows by various modes, and so forth, and those factors that affect these flows. In general, movement within a single industrial firm or building, or the migration of population, is not included in this area." [p13, 20]

The essence of transportation is not planes, trains, and automobiles, but rather **mobility** and **access** [p3, 21]. The interaction between travellers and environment is emphasised when studying mobility in transport systems. This is the core of **transport geography**, "a sub-discipline of geography concerned about movements of freight, people and information. It seeks to understand their spatial organisation by linking spatial constraints and attributes with the origin, the destination, the extent, the nature and the purpose of movements."[p6, 22]

An essential component of transport systems is a variety of modes by which people can move in space and time. The passenger sector provides various modes for selection: walk & bike, bus, passenger rail, aviation, light-duty vehicle, and 2-wheel or 3-wheel vehicles. Another common taxonomy used in urban mobility is public transit (PT) and car, be it privately owned or shared among users. PT is a system of transport for moving large groups of passengers, which is available for the general public. It covers multiple sub-modes such as bus, train, subway, railway, and tram etc. PT offers scheduled services in rigid networks where travellers need to adapt their plans accordingly [23]. Ride-sourcing is an emerging car-based mode, with rapid

growth worldwide in the use of phone-based ride-hailing applications such as Uber, Lyft, and DiDi Chuxing. As opposed to private car, ride-sourcing represents a trend of shared mobility i.e., the services and resources involved in using a motor vehicle, bicycle, or other low-speed transportation mode that is shared among users, either concurrently or one after another [24].

Transport mode is a key determinant to the emissions and it contributes to 30% of world greenhouse gas (GHG) emissions. According to the Fifth Assessment Report of the United Nations Intergovernmental Panel on Climate Change, after 2030, transport is likely to become the sector with the highest emissions in the 2°C scenario. Different modes have distinct characteristics such as load factor (number of passengers/capacity per vehicle) and carbon intensity (fuel economy), therefore contributing to the overall carbon emissions differently. A recent study suggests that occupancy explains about 70-80% of the variation in the GHG intensity of major passenger transportation modes [25], therefore, a more sustainable mix of transport modes becomes increasingly important.

Besides increased GHG emissions, increased car use worldwide especially in developing countries has many other negative environmental impacts, including traffic congestion, land-use issues such as parking, and increased air pollution. PT can provide a low-cost, energy-efficient, less polluting, and socially equitable travel alternative [26, 27]. Policymakers worldwide recognise the importance of promoting a mode shift from car to PT and other low-carbon modes in cities as a way to address negative environmental impacts, increase equity [28], and combat climate change [29].

While there is strong evidence that car-based travel is often faster than public transit, the spatial and temporal patterns of this time discrepancy are crucial to better inform urban planning and policy efforts to encourage travel mode shifts. On the other hand, the potential of ride-sourcing services to replace PT trips has largely been overlooked [30], which might cause increased carbon emissions. Despite some efforts, the relationship between ride-sourcing and PT remains elusive. These point to the importance of a better understanding of the modal disparities between car and PT at high granularity, in order to encourage the mode shifts from car-based travelling to more use of PT.

2.3 Data sources

2.3.1 Mobility traces

In the last decade, the emerging data sources have significantly improved our understanding of mobility [8, 15, 31]. Common emerging data sources are call detail records (CDR), tracking apps on smart phones, GPS-enabled devices, and geotagged social media.

The data sources of mobility traces have two forms: longitudinal and

lateral. A **longitudinal** dataset is characterised by the long-term (more than 24 hours) and continuous observations focusing on a group of participants, such as GPS log [e.g., 32], app-based GPS log data [e.g., 33], CDR [34], and Twitter users' geotagged activity trajectories [e.g., 35]. Longitudinal datasets are often applied to reveal the patterns of individual mobility, e.g., the socio-geography of mobility [36] and the activity space estimation [37]. Because it is possible to observe the individual trajectory over a long period of time, more attention has been paid to the routine mobility [38] and next-location prediction [39]. A **lateral** dataset is often collected based on a particular area, such as a city or a country, during a short-to-medium period, and it usually covers a larger population. It is commonly used to study the travel demand [40] and behaviour patterns at the population level [41]. The difference between the aforementioned two data forms is due to the practical trade-off between the number of individuals and data collection duration; that is, for the longitudinal form, the data collection duration is short but the covered number of individuals is limited, while the lateral form can cover a much larger population, but the time needed for data collection is much longer.

Here five data sources are discussed in detail: household travel surveys, CDR, GPS log data, App-based GPS log data, and social media data. The main characteristics of the five data sources are summarised briefly in Table 2.1 based on the literature review presented in the upcoming subsections. Compared with the other data sources, social media data have strengths in long collection duration, a large number of studied individuals, large spatial coverage, ease of access, low cost, and accurate location information. The main weaknesses are the incomplete sampling of individual trajectories and lack of socio-demographic information and trip information such as trip purpose and travel mode.

Table 2.1: Characteristics of the five data sources. [a] Geotagged social media data. [b] Traditional household travel survey. [c] Time length of tracking the same individual. [d] Low cost = +++. Medium cost = ++. High cost = +.

	Check-ins[a]	Travel survey[b]	CDR	GPS log	App-based GPS log
Time duration[c]	+++	+	+++	++	+++
Number of individuals	++	+++	+++	+	++
Spatial coverage	+++	++	++	+	+++
Trajectory completion	+	+++	++	+++	+++
Accessibility	+++	++	+	+	+
Cost[d]	+++	+	++	++	++
Spatial resolution	+++	++	++	+++	+++
Temporal resolution	+	+++	++	+++	+++
Socio-demographic info.	✗	✓	✗	✓	✗/✓
Trip info.	✗	✓	✗	✗/✓	✗/✓
Passive collection	✓	✗	✓	✗	✓

Household travel survey

Due to the lack of longitudinal data, most previous studies used lateral data [42] among which household travel surveys were the most prevalent. Pucher et al. [43] analysed the 2001 and 2009 National Hoursehold Travel Surveys to understanding how the daily walking and biking behaviour changes at the population level. Liang et al. [44] revealed the exponential law of intro-urban mobility based on a one-year of 46, 000 trips between 2017 zones within a county.

Travel surveys contain socio-demographic information and detailed activity records making them not easily replaceable by other emerging data sources [45]. Because their sampling is carefully designed to derive statistically representative population-level estimates, traditional travel surveys remain a vital source for validation/calibration of the emerging data sources. But they also have many shortcomings such as being costly to collect and having low sampling rates, short survey duration, under-reporting of trips, and quickly being out-of-date [46]. Travel surveys also fail to capture most of the long-distance trips [45].

Mobile phone CDR

Mobile phone CDR are the most widely applied among these emerging data sources [7]. A record in a CDR dataset represents a phone call or a text message with the phone activity information (start time, duration, and end time, etc.) and the GPS coordinates of the tower that first channelled the activity. This implies that the spatial accuracy of an individual location depends on the cell tower network's spatial coverage, typically 200-300 meters. From the perspective of individual trajectory, Phithakkitnukoon, Smoreda and Olivier [36] explored geo-social radius of individuals using one year of anonymised call detail records of over one million mobile phone users in a country; in order to identify the privacy bounds of human mobility, De Montjoye et al. [47] collected data from 1.5 million users of a mobile phone operator in a country for one year. In addition, the application of CDR has matured for understanding the clustering structure of spatial interactions [48] and developing OD matrices [49].

CDR can be collected long-term with very large numbers of tracked individuals. For example, a study uses one-year-long CDR series with nearly 15 million tracked individuals to study the impact of mobility on malaria [34]. Nevertheless, this data source is often not easy to access, and, compared with travel surveys, has the shortcomings of spatiotemporal sparsity and incomplete trajectories [50]. It is also often not available for follow-up tracking and continuous update.

GPS log data

GPS log data contain the records of GPS coordinates sampled in the frequency that is regular and high (e.g., one log per 10 seconds [32]). Applied GPS log data can be divided into two main categories: human-carried GPS logger and vehicle-attached logger. The latter is beyond the scope of this thesis. Rhee et al. [51] revealed the Levy-walk nature of human mobility based on 101 individuals' GPS traces collected in five outdoor sites over 226 days. De Domenico, Lima and Musolesi [52] explored the predictability of human mobility and social interactions using a dataset collected from 25 individuals over one year in a country. A large amount of studies seek the good performance of individuals' future location prediction [e.g., 53].

Most previous studies apply GPS log data from a rather small group of individuals (20-500). Most of these studies come from the computer science community focusing on the individual-based prediction of future locations [e.g., 54]. Compared with CDR and household travel surveys, such a data source is used less frequently by the transport research community due to small sample size, high cost, and lack of modal travel information (even though some research efforts specialise in deriving modal estimates from the logged data [e.g., 55]). Overall, GPS log data provide a relatively complete and accurate picture of individual mobility trajectory, making it close to the "ground truth."

App-based GPS log data

Recently, GPS locations have been collected through the use of apps such as tracking apps or activity trackers installed in mobile phones, tablets, or smartwatches [56]. This has significantly increased the scope of data collection to an unprecedented level across space, time and users. For example, [33] applied data from 700,000 users with high-resolution traces using smartphone apps spanning three years. These tracking apps on smartphones provide data of high spatiotemporal resolution, long-term observation of individuals, and self-reported socio-demographics such as age and gender. However, these data are costly to collect and often have limited access due to privacy concerns.

Compared with small-scale GPS log data, app-based GPS log data can cover a larger population and a longer data collection time period. Compared with CDR and household travel surveys, such a data source features long-time data collection, limited data accessibility, and lack of modal travel information. And their population coverage is often smaller than CDR. However, the smartphone-based prompted recall travel survey is gaining attention which aims to support passive GPS data collection with user-reported travel mode and activity information [56]. Increasingly popular app-based GPS log data outperforms the conventional GPS log data on its completeness and accuracy bringing the "ground truth" of human mobility even closer.

Social media data

In this thesis, we use Twitter as the representative of social media data. A tweet typically contains multiple components that can be useful for transport research, including text, hashtag, location, and timestamp. When users choose to have their location reported when sending out tweets, these are called **geotagged tweets**. Geotagged tweets account for a small proportion (1-3%) [57]. That number varies between regions, 7.4% (George, South Africa), 1.9% (Barcelona, Spain), 1.1% (Kuwait), and 0.3% (Sweden) [58]. Despite the low proportion of geotagged tweets, these check-ins provide precise location information and have increasingly been used for understanding mobility [59, 60].

Geotagged tweets can be obtained in three ways: 1) Purchase the complete set of public tweets from Twitter Firehose; 2) Access the Streaming API to get a maximum of 1% of the public tweets; 3) Access the user timeline by user name/ID to get a maximum of 3200 historical tweets that are set by the user as publicly accessible.

Geotagged tweets collected from the Streaming API are often limited to a geographical bounding box yielding a lateral dataset. It covers a large number of Twitter users but takes time to accumulate enough samples, and individuals' movements across the bounding box are not captured [10]. Most studies use geotagged tweets in this form, i.e., focusing on a specified area that is often in line with the spatial scale of policy-making and urban planning. For lateral data, the individual trajectory of geotagged tweets is often aimed at validation and understanding of fundamental laws of human mobility, such as the power law distribution of trip distance [60]. Compared to individual trajectories, the perspective of places networks gains more attention because they connect directly to travel demand modelling and have greater potential to support applications such as modifying the classic gravity model by integrating locations posted on Foursquare [40]. Gao et al. [61] validated OD trips mined from the geotagged tweets against the large-scale studies' results using more than 6 million geotagged tweets collected over one month.

By accessing the user timeline, all the publicly available historical tweets by a specified user can be collected resulting in a longitudinal record of the individual trajectory without any geographical boundaries. Longitudinal geotagged tweets are the only data source that is not constrained to a specific area. This type of longitudinal data has been scaled up to large numbers of Twitter users to study the influence of global cities on human diffusion [62]. Hasnat and Hasan [63] used geotagged tweets to identify tourists and to study the spatial patterns of their destinations. Exploring urban mobility and neighbourhood isolation, Wang et al. [4] analysed 650 million geocoded Twitter messages to estimate the home locations and travel patterns of almost 400,000 residents in 50 largest cities in America over 18 months.

The low cost of retrieving geotagged tweets makes them especially appealing compared to other data sources [9]. The data source is free to access, and

it provides precise location information with a spatial resolution of around 10 meters compared with 100-200 meters for call detail records (CDR) [60]. Moreover, it allows for long-term tracking of movements that are free of geographical boundaries [35].

The main criticism pertains to two aspects, a **biased population representation** and **low and irregular sampling**. There have been studies comparing multiple data sources to identify/adjust the biases [e.g., 64, 65] and to validate against "ground truth" [e.g., 59]. When validating geotagged tweets against travel surveys, one study shows that geotagged social media data capture the displacement distribution, length, duration, and start time of trips reasonably well for inferring individual travel behaviour [66]. Validations using CDR need to be interpreted carefully as CDR and geotagged tweets have similar passive data collection manners that might share some shortcomings. Some studies have compared geotagged tweets with traffic data [67] and travel-demand data [68], generally achieving good results.

Despite the known disadvantages of geotagged tweets, one recent literature review shows that experts are positive about the usefulness of such data sources for modeling travel behaviour [9]. There is also a consensus on the need for careful inspection of using geotagged social media data to approximate the actual travel behaviour of the general population.

2.3.2 Geographical Information Systems (GIS) data

GIS refers to "a set of powerful tools for collecting, storing, retrieving at will, transforming, and displaying spatial data from the real world for a particular set of purposes" [p3, 69]. As shown in Figure 2.1, there are three main feature classes in GIS for transport: **transportation network**, **population flows**, and **land use patterns**; the four major components, encoding, management, analysis, and reporting have their specific considerations for transportation.

Rapidly increasing amount of data sensed in urban transport systems from GIS [70, 71] have deepened our understanding of different modes in the transport systems. The data sources used in this thesis include open data for road networks [72], public transit schedules and routes [73], traffic speed data [74], and open trip data for ride-sourcing.

Road network

Transport networks of various modes are represented as a set of interconnected lines, such as roads and rail lines, making up a set of features through which individuals can flow [p214, 75]. A network graph defines potential movements from node (place) to node including prohibited and permitted

[1] Adapted from The Geography of Transport Systems: `https://transportgeography.org/?page_id=6578`

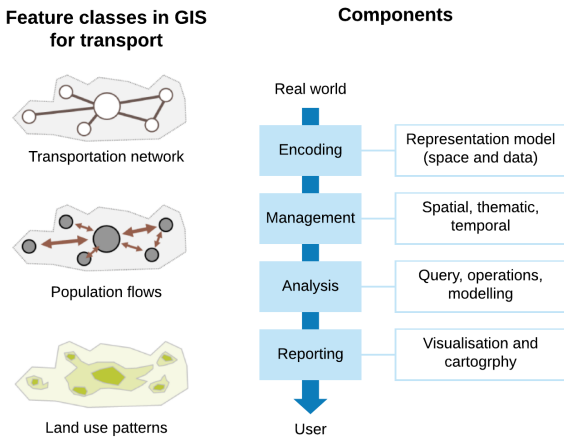


Figure 2.1: Components of GIS and major classes for transportation.[1]

connections and the possible direction of movement on a link in terms of whether it is one-way in a particular direction or bidirectional [p339, 20]. Transport networks including rich attributes e.g., distance and speed limit of each network link are available via **OpenStreetMap** [76], a collaborative project to create a free editable map of the world. An example of downloaded street network of Modena, Italy is shown in Figure 2.2 using a Python package, osmnx [77].

General transit feed specification (GTFS)

General Transit Feed Specification (GTFS) is one of the open data standards for public transit proposed by Google. A GTFS static dataset [78] is a collection of text files consisting of all the information required to reproduce a transit agency's schedule, including the locations of stops and timing of all routes and vehicle trips. GTFS data can be collected from various sources that are publicly available. For example, in this thesis, some GTFS data were obtained from OpenMobilityData [79] uploaded by local agencies. Figure 2.3 shows an example of PT lines contained in a GTFS dataset from Stockholm.

Traffic data

The availability of real-time traffic speed data enables more advanced traveller information systems for route choice and better-informed traffic planning [80]. Emerging data sources, such as **HERE Traffic** [74] with extensive coverage of cities in 83 countries to date [80], can collect and provide information about real-time road speed, incidents, and accidents. The amount of available data and the level of spatial and temporal details allow more realistic

Figure 2.2: OSMnx street networks automatically downloaded and visulised for Modena, Italy. Adapted from the source: Figure 4 in [77].

estimates of travel time and congestion level [81].

Mode-specific trip data

In order to understand the transport modal disparities in urban mobility, collecting trip data at large spatiotemporal scales is important. Conventional surveys are limited by their small sample sizes, deviations from actual travel behaviours, and failures of incorporating the built environment.

The recent development of GPS-enabled devices allows for fast accumulation of a massive amount of spatial data, offering new opportunities. For example, the City of New York has an open data portal for taxi trips [82]. Many studies have used these trip data to answer a variety of questions [83–86] including modelling taxi demand in New York City [83] and the impact of time and weather on taxi ridership [84] using a dataset of 147 million taxi-trip records covering 10 months. Big data analytic tools are used to explore the factors that motivate massive amounts of trips by transit, taxi, and bike-sharing in Washington, D.C [30]. Using big trip data, the relationship between taxi and transit are divided in three categories: transit-competing, transit-complementing, and transit-extending [86].

Increasing amount of mode-specific data are made freely available to the public. However, coverage of a large area and population is often achieved

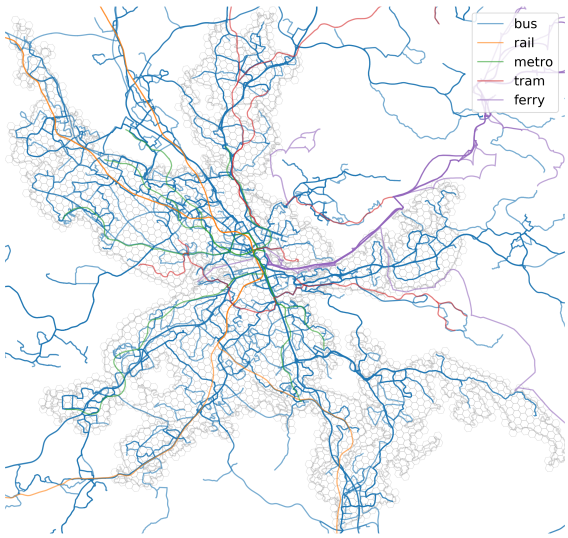


Figure 2.3: PT lines in Stockholm. Source: Liao and Gil.

at the cost of rich detail, such as trip purpose, compared to conventional survey-based data. Open data often only contain the geolocation of origin/destination and partial trajectories without trip purpose. To make full use of the data requires data enrichment where external data sources are often needed. For instance, in order to reveal the shared-use mobility competition at the trip level, data from various sources are combined [30] including taxi trips, metro line trips, census, and OpenStreetMap [76]. The latter provides crowd-sourced built environment characteristics and transportation network connectivity for a better explanation of the observed trip patterns.

CHAPTER 3

Methodology

This chapter starts with an overview of data science, as the core of the methodology framework of this thesis. Specifically, three categories of methods are introduced in this chapter: data mining (Section 3.1), mobility metrics and models (Section 3.2), and methods in transport geography (Section 3.3). Each section describes the particular methods applied in the appended papers.

This thesis is organised surrounding a keyword, data, in the context of understanding human mobility and transport modal disparities. The emerging data sources introduced in Chapter 2 are attributed to the prevalence of digital technologies permeating into every aspect of modern life. Unprecedentedly, human activities and natural records that occur in the whole planet are more and more registered. The term "big data" became widespread as recent as 2011 [87]. Oftentimes people ask how large a dataset is qualified to be called as "big data"? The volume is just part of the story. The term "big data" also highlights the use of advanced data analysis methods that extract value from data [88], where the traditional techniques fail to work efficiently or effectively.

When people are hyping "big data", data itself is often overly emphasised causing the impression that bigger data naturally bring deeper insights. These large amounts of data create an unprecedented situation where we think more of: "let me play with data to see what I can get from them." Suddenly, a hammer called "big data" is handed over to us and we start searching nails everywhere. However, we should always ask: "I have this question, what data do I need?" Without the right questions and methods, data are just data.

The role of data science in this big data world is like the importance of oil refinery for crude oil [p1, 89]. Data science is a multi-disciplinary field that intersects between Computer Science/Information Technology, Mathematics and Statistics, and Domains/Business Knowledge. This thesis sits in data science for leveraging new data sources to contribute to the domain knowledge of mobility and transport geography. The methodological framework is shown in Figure 3.1.

The intersected between Computer Science/IT and Mathematics & Statistics is Machine learning under which Data mining (Section 3.1) is applied in

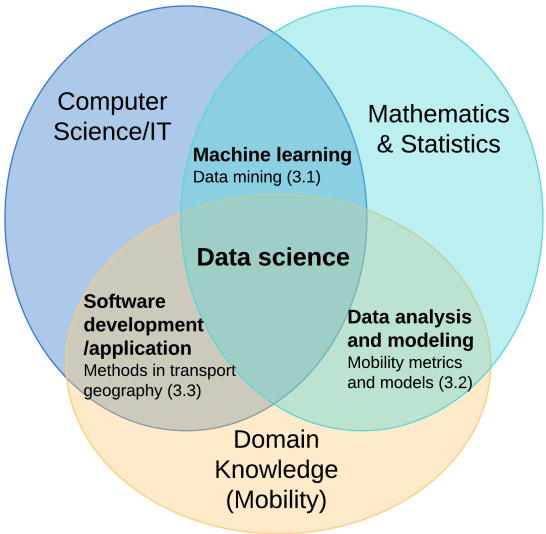


Figure 3.1: Methodology of this thesis. Appended papers apply different methods that are introduced in this chapter.

Paper I to reveal the population heterogeneity of mobility using Twitter data and in Paper V to explore the relationship between ride-sourcing trips and their PT alternatives. The intersected part between Mathematics & Statistics and Domain knowledge of mobility represent the traditional data analysis and modelling where the general mobility metrics and models (Section 3.2) are shared by all the appended papers, particularly in Paper II and III about the travel demand modelling. In Paper IV and V, the methods in transport geography (Section 3.3) applied lie in the inter-discipline of Computer Science/IT and domain knowledge of mobility; they are used to calculate the travel time of using car and taking PT in a data-driven manner as well as the spatial analysis of ride-sourcing trips and their PT alternatives. The usage of different methods are summarised in Table 3.1.

Table 3.1: Methods applied by the appended papers.

Section	Methods	Paper				
		I	II	III	IV	V
3.1	Data mining	✓				✓
3.2	Mobility metrics and models	✓	✓	✓	✓	✓
3.3	Methods in transport geography				✓	✓

3.1 Data mining

Big data in mobility imposes new challenges such as a large scale, a high complexity, and privacy sensitivity. Therefore, it requires cutting-edge research and development where recent advances in machine learning (ML) provide a vast set of tools that can analyse mobility data [90], but choosing the right tool for a given task is vital. A detailed review can be found in the survey paper by Toch et al. [90].

Data mining itself is a multi-disciplinary field under or sometimes overlapped with ML. It is an iterative process within which progress is defined by predictive or descriptive discovery, through either automatic or manual methods, and it is most useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome [p2, 91]. There are many data-mining techniques, such as regression, classification, and clustering. Unlike some other ML techniques, such as deep learning, are a less interpretable black box, the success of a data-mining engagement depends largely on the amount of energy, knowledge, and creativity that the designer puts into it [p3, 91]. It emphasises the importance of domain knowledge and the interpretable results which make it a particularly powerful tool for obtaining knowledge of human mobility. A common data-mining process is shown in Figure 3.2.

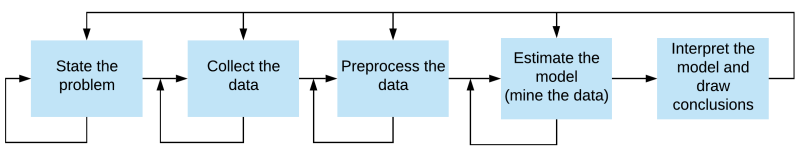


Figure 3.2: The data-mining process. Adapted from Figure 1.2 in [91].

When learning from data with the estimated model, there are two types of inductive-learning methods; unsupervised learning such as cluster analysis, and supervised learning such as classification. The rest of this chapter introduces the particular part of data mining that has been applied in Paper I and V. For further information, a comprehensive description of data mining can be found in the book by M. Kantardzic, 2011 [91].

3.1.1 Cluster analysis

As one essential part of data mining, **cluster analysis** consists of a series of methods for automatic classification of samples into a number of groups using a measure of association so that the samples in one group are similar and samples belonging to different groups are not similar [p250, 91]. **The input to a cluster analysis is described as a series of feature sets that are**

normalised first, $F_i = [f_1, f_2, ..., f_n], i = 1, 2, ..., N$ where we have N samples that are to be classified. Without pre-defining how many classes we expect, we propose n features to describe the object based on domain knowledge. For example, the mobility metrics such as trip distance are used in Paper I as one of the features describing the individual mobility trajectory. **Output from the clustering analysis is a partition** $\Lambda = \{G_1, G_2, ..., G_K\}$, where $G_k, k = 1, 2, ..., K$ is a crisp subset of the input samples such that

$$\begin{aligned} G_1 \cup G_2 \cup, ..., \cup G_K = F_1, F_2, ..., F_N \text{ and} \\ G_{k1} \cap G_{k2} = \emptyset \text{ for } k1 \neq k2. \end{aligned} \tag{3.1}$$

And the members of Λ are called clusters. Sensible clustering is measured by the small sum of squares of deviations within the same cluster. By limiting the cluster distance larger than a certain threshold, the final clusters are formulated. The average silhouette width provides an evaluation of clustering validity [92]. There are two categories of cluster analysis; hierarchical clustering, and iterative square-error partitional clustering e.g., K-means method.

Hierarchical clustering

Hierarchical techniques organise data in a nested sequence of groups, which can be displayed in the form of a dendrogram or a tree structure [p252, 91]. A two-dimensional illustration of hierarchical clustering is presented in Figure 3.3. This method constructs a binary tree of the data that consecutively combines samples that are close in terms of certain similarity measures. Cutting the similarity tree by certain criteria gets you a different number of clusters.

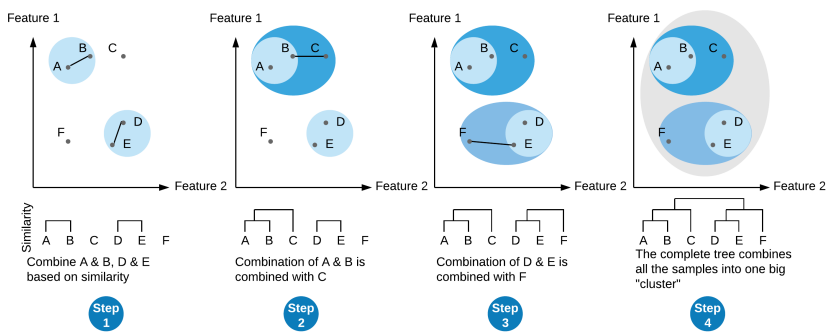


Figure 3.3: A two-dimensional example of Hierarchical Clustering.[1]A-F are samples that are described by Feature 1 and Feature 2. The similarity is measured by the distance between samples on the chart. Closest samples are combined first.

A general process of hierarchical clustering is illustrated in Table 3.2. Feature construction is using domain knowledge to select important features to describe the study object. Step 2 is necessary for calculating the distance to avoid the effect of the unit which otherwise over-weights those features with large values (100 m will be weighted more than 0.1 km). The step of distance calculation is to measure the similarity between samples' feature sets. The squared Euclidean distance [93], widely adopted in previous studies, is applied in Paper I. To establish cluster linkages, Ward's method was used where the decrease invariance for the cluster being merged [94].

In Paper I, using hierarchical clustering, each Twitter user/traveller is categorised into a group with certain mobility patterns where four groups are constructed with their distinct mobility patterns.

Table 3.2: Procedure of Hierarchical Clustering.

#	Step	Paper I
1	Feature construction	Mobility metrics
2	Data normalisation	Max-min normalisation
3	Distance calculation	Squared Euclidean distance
4	Linkage establishment	Ward's method
5	Split linkage into clusters	Similarity threshold
6	Cluster structure evaluation	Silhouette Width

K-means clustering

K-means clustering is a partitional algorithm which produces clusters by optimising the square-error criterion. The objective is to obtain the partition that the squared error between the empirical mean of a cluster and the vectors in the cluster is minimised [95].

The algorithm starts with a selection of an initial partition of K clusters that contain randomly chosen samples, the centroids of the clusters are calculated as $\mu_k = (1/n_k)\sum_{i=1}^{n_k} x_{i,k}$, where $x_{i,k}$ is the sample i of cluster G_k. Next, a new partition is generated by assigning each sample to the closest cluster centre. The square-error of a new cluster G_k i.e., the within-cluster variation is the sum of the squared Euclidean distances between each sample in G_k and its centroid μ_k, $e_k^2 = \sum_{i=1}^{n_k} \parallel x_i - \mu_j \parallel^2$. The overall square-error is the sum of all the clusters' within-cluster variation, $E_k^2 = \sum_{k=1}^{K} e_k^2$. With the new clusters, update the centroids and repeat the process until an optimum value of E_k^2 is found.

In Paper V, K-means clustering is applied to form functional regions based

[1] Adapted from BRANDIDEA: `https://www.brandidea.com/hierarchicalclustering.html`

on the point of interest (POI) profile of the study area, where the ride-sourcing trip records generated, characterising the built environment of where the trips were originated from and attracted to.

3.1.2 Classification

Classification is a type of data analysis that creates models i.e., classifiers describing data of pre-defined categorical classes. For example, we can build a classifier based on a set of labelled pictures of cat and dog and such a classifier can be applied to predict a new picture having a dog or a cat. Unlike unsupervised clustering techniques, classification deals with labelled data and attempts to distinguish the classes. The goal of this supervised learning technique is to learn a predictive model that maps features of the data (e.g. hair length, location of eyes, ear size, ...) to an output (e.g. cat or dog).

Using machine learning techniques, such as classification models, raises an issue of interpretability: insights about the data and the task the machine solves are hidden in increasingly complex models [ch1.2, 96]. If accuracy is the only target, more and more complex models will win the game. However, in most cases, we care more than just accuracy about prediction. We want to understand why the model performs in certain ways and what insights we can learn from the feature space that distinguishes the interested classes the best. That's why the concept of **interpretable machine learning** is gaining increasing attention.

Common interpretable models are linear regression and its extensions, logistic regression, decision trees etc, which people can easily understand and interpret. These models hold potentials for better synthesising the various dimensions of observations and understanding the differences between classes. For instance, in the generalised additive model (GAM) [97], a generalised linear model, the linear part of the variable depends linearly on unknown smooth functions of the independent variables. Model construction focuses on the inferences about these smooth functions. The recent machine-learning techniques have enhanced the traditional GAM by bagging, gradient boosting, and automatic interaction detection [98]. Compared with classic glass-box models such as logit models, this **enhanced GAM** generally delivers more accurate results, while keeping them insightful and easy to visualise.

The enhanced GAM originates from the traditional Generalised Additive Model (GAM) [97]:

$$g\left(E\left[y\right]\right) = \beta_0 + \sum f_i(x_i) \tag{3.2}$$

where g is a link function connecting the expected value of y with the right part of the equation, β_0 is a constant, and $f_i(x_i)$ is an unknown smooth function of x_i. The logit function is a common link function for binary clas-

sification. GAM has subsequently been modified into a model called GA^{2}M [99] that allows interactions between explanatory variables to be captured:

$$g\left(E\left[y\right]\right)=\beta_0+\sum f_i\left(x_i\right)+\sum f_{ij}\left(x_i,x_j\right) \tag{3.3}$$

This increases accuracy while keeping a high level of intelligibility. The training process of GA^{2}M finds the form of variable smooth functions. GA^{2}M is further enhanced by modern-machine learning techniques to train GA^{2}M faster while allowing for large datasets [98]. It also enables automatic interaction detection.

In Paper V, we define a ride-sourcing trip as being transit-competing or non-transit-competing trips. This enhanced GAM is used to characterise the two categories of ride-sourcing trips to better understand the relationship between public transit and ride-sourcing.

3.2 Mobility metrics and models

In physics and mathematics, there are fundamental metrics used to characterise mobility as it is a process of the geographic displacement of human beings, seen as individuals or groups, in space and time. This displacement constitutes of an origin, a destination, and a specific trajectory in between (Section 2.1). The corresponding metrics and models are summarised in Figure 3.4.

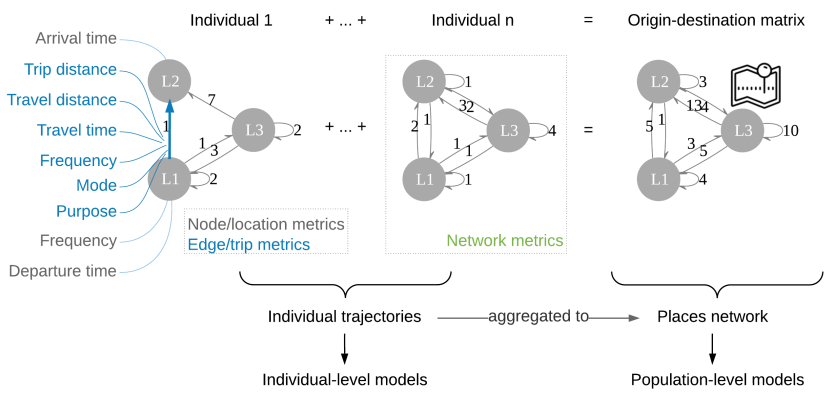


Figure 3.4: A framework of mobility metrics and models. L1-3 are three distinct locations/zones. The edges/arrows pointing from one location to another are trips that connect an origin and a destination. The numbers next to the edges are the frequency of the observed trips based on the individual trajectory or the aggregated origin-destination matrix.

3.2.1 Mobility metrics

If we can track any given individual continuously, his/her location trajectory can be expressed as a series of locations with time stamps: $\mathbf{L}_p = (X, Y, t)_{p,k}$, $k = 1, 2, ..., N_p$ where X is the decimal degree of Latitude, Y is the decimal degree of Longitude, t the time stamp (UTC) of the k-th location. The number of distinct locations is smaller than the total number of locations he/she visited. Let n_p be the number of distinct locations and $\mathbf{T}_{p,i}$ be the series of times when visiting location i either as an origin or a destination. The vector of visited distinct locations is therefore:

$$\mathbf{L}'_p = (X, Y, \mathbf{T})_{p,i}, i = 1, 2, \ldots, n_p \tag{3.4}$$

where $\mathbf{L}'_p$ formulates a complete network of distinct locations. One realisation of an edge in this network is called a **trip: the connection between two consecutive stays generated by the same individual** (p).

A trip can be characterised by many indicators. **Trip distance** $(d_{i,j})$ refers to the Haversine distance between the origin (i) and the destination (j) where the Haversine formula is used to calculate the great-circle distance between two points. This distance is the shortest distance over the earth's surface. It is similar to the straight line distance when the two locations are close to each other. However, when the two locations become far away from each other so that the earth's surface is not neglectable, the straight line distance does not fit anymore. **Travel distance** $(D_{i,j})$ refers to the actual distance/network distance by summing up the travelling trajectory given fine enough sampling resolution. **Travel time** $(TT_{i,j})$ is the time spent from one location to reach another location by a certain **mode** $(m_{i,j})$. Travel time is roughly proportional to the distance travelled given a certain mode of transport, which itself depends on the trip distance. For short-range travel, slow modes e.g., walking and public transit with many stops are used, while for longer distances, one typically takes fast trains or planes with comparatively fewer stops [15]. **Trip frequency** $(f_{i,j})$ refers to how frequently trips are formulated between two locations. **Trip purpose** $(P_{i,j})$ refers to the purpose of this trip, e.g., work and leisure. For example, usually the connection between workplace and home has much higher frequency than the other location pairs.

Considering the above fundamental metrics, the mobility trajectory of the individual p formulates a network of distinct locations $(\mathbf{G}_p)$.

$$\mathbf{G}_p = (d, D, TT, m, f, P)_{i,j}, i, j = 1, 2, \ldots, n_p \tag{3.5}$$

And aggregating $\mathbf{G}_p$ through Individual $p = 1$ to Individual n for all purposes gives the movement flows of population formulating a network of places (see Figure 3.3). It is also called an **origin-destination (OD) matrix** in mobility studies and transport planning which has the below basic form

$$\mathbf{G} = (d, F)_{i,j}, i, j = 1, 2, \ldots, N \tag{3.6}$$

where $F_{i,j}$ is the total number of individuals travelling between zone i and zone j. And N refers to the total number of distinct locations/zones.

Refocusing to locations, **location frequency** represents how frequently it is visited either as an origin or a destination. The series of times when visiting location i, $\mathbf{T}_{p,i}$, provides a temporal profile with this location. This temporal profile is a crucial representation of human mobility (see Figure 3.5). At the individual level, it tells one's lifestyle and it helps to predict one's mobility. At the aggregate level, this metric helps to capture the "heartbeat" of a city.

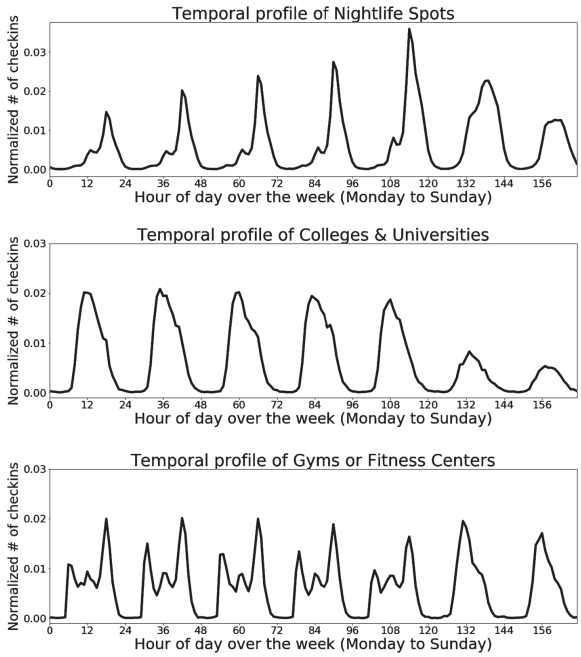


Figure 3.5: Distinct temporal profiles of different venues. Source: Figure 2 from [100].

At the individual level, the diffusive behaviour of humans at certain scales suggests that they tend to move a characteristic distance away from their starting locations [15]. This distance can be quantified by an important construct, **radius of gyration** (r_g). It refers to the travel distance range weighted by the visiting frequency. The total radius of gyration r_g is defined as:

$$r_g = \sqrt{\frac{1}{n_p} \sum_{i=1}^{n_p} f_i \cdot (\mathbf{r}_i - \mathbf{r}_{cm})^2} \tag{3.7}$$

where $\mathbf{r}_i = [X, Y]_i$ and the mass centre of the visited locations:

$$\mathbf{r}_{cm} = \left[\frac{\sum_{i=1}^{n_p}(X_i \cdot f_i)}{\sum_{i=1}^{n_p} X_i}, \frac{\sum_{i=1}^{n_p}(Y_i \cdot f_i)}{\sum_{i=1}^{n_p} Y_i} \right] \tag{3.8}$$

There are various network metrics to describe the structure of $\mathbf{G}_p$ which are also applicable to the aggregated OD matrix. Here, a few network metrics are selected to present at the individual level as they are used in Paper A. **Clustering coefficient (average)**, $\overline{C}$ (-), refers to the degree to which the neighbours of a given node link to each other [p63, 101]. For a node (location) i with degree (visiting frequency) $f_{p,i}$, its local clustering coefficient is defined as:

$$C_i = \frac{2L_i}{f_i(f_i - 1)} \tag{3.9}$$

where L_i indicates the number of links between the k_i neighbours of node i. The average clustering coefficient of the whole network is calculated by:

$$\overline{C} = \frac{1}{n_p} \sum_{i=1}^{n_p} C_i \tag{3.10}$$

The mean value of the log-transformed node degree, z (-), represents the overall visiting frequency. Each visited location is seen as one node in the network, and the visiting frequency is equivalent to the node degree; therefore, the average value of the node degree z is one important indicator of the network properties. It is defined as:

$$z = \frac{\sum_{i=1}^{n_p} \log(f_i)}{n_p} \tag{3.11}$$

z_m (-) is **the max node degree divided by the sum of total degrees**, which indicates the how centralised the overall visited locations are. The normalised max node degree z_m is defined as:

$$z_m = \frac{\max[f_i]}{\sum_{i=1}^{n_p} f_i} \tag{3.12}$$

These metrics constitute the essential building blocks for the understanding of how people move in space and time. They have been widely used in the literature for reproducing individual mobility patterns or general population flows to reveal spatiotemporal patterns of mobility with models. The rest of this section dives into the models that build on the metrics.

3.2.2 Statistical models

In order to understand the mobility patterns, some studies have been focused on the statistical characterisation of trip distance distribution since the dawn

of the big data era, when a massive amount of traces data started becoming available, such as the circulation records of banknotes [102] and the call detail records [31].

Using these unprecedented data sources, the power-law paradigm has been the most popular way to quantify the trip distance distribution [31, 103]. However, many studies have argued that human mobility is not always scale-free depending on the spatial scale [33] and transport mode [104], where people found other functions such as Weibull and lognormal are more suitable [105, 106]. Common models used for characterising the probability density function $f(d)$ of trip distance d are summarised in Table 3.3.

Table 3.3: Common probability density functions of trip distances $f(d)$ ($d > 0$).

Model	Equation	Parameters
Weibull	$\frac{k}{\lambda}\left(\frac{k}{\lambda}\right)^{k-1} e^{-(d/\lambda)^k}$	k and λ
Gamma	$\frac{\beta^\alpha d^{\alpha-1} e^{-\beta d}}{\Gamma(\alpha)}$	α and β
Exponential	$\lambda e^{-\lambda d}$	λ
Truncated power law	$(d+d_0)^{-\beta}\exp(-d/K)$	d_0, β, and K
Power law	$(d+d_0)^{-\beta}$	d_0 and β
Lognormal	$\frac{1}{d}\cdot\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln d-\mu)^2}{2\sigma^2}\right)$	μ and σ

Their distributions are shown in Figure 3.6. In Paper III, we model the trip distance distributions using these theoretical models for the model-synthesised mobility trips.

3.2.3 Individual-level models

To some degree, individual mobility can be regarded as uncertain because of arbitrariness in the actions of individuals, leading to a certain level of stochasticity. However, individual trajectories are far from random in reality, displaying a high degree of regularity and predictability, which can be exploited to predict an individual's future whereabouts and to construct realistic generative models of individual mobility [15].

The basic models reproducing individual mobility are called random walks in the discipline of Complex Systems. The location of individual p, $\mathbf{L}$ starting from $(0,0)$, after N_p steps of movement becomes

$$\mathbf{L}\left(t_{n_p}\right) = \sum_{i=1}^{N_p} \Delta\mathbf{L}(t_i) \tag{3.13}$$

where $\Delta\mathbf{L}(t_i)$ is the jump on time t_i which is a random variable from a probability distribution $f(\Delta\mathbf{L})$. And jumps are assumed to be statistically independent.

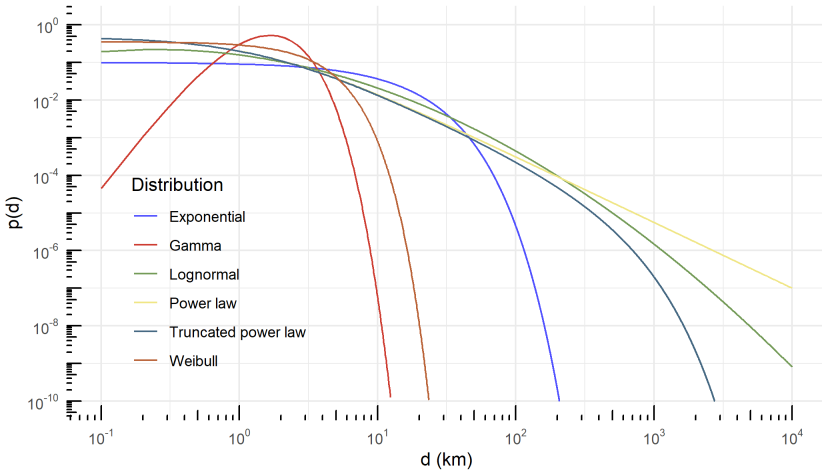


Figure 3.6: Statistical models for probabilistic density function of trip distance. Weibull - $k = 1.5, \lambda = 3$, Gamma - $\alpha = 6, \beta = 3$, Exponential - $\lambda = 0.1$, Truncated power law - $\beta = 1.8, K = 300, d_0 = 1.5$, Power law - $\beta = 1.8, d_0 = 1.5$, and Lognormal - $\mu = 1.5, \sigma = 1.7$.

The scaling of the square root of the mean squared displacement (RMSD) is particularly interesting for studying individual mobility:

$$R(t) = \sqrt{\langle \mathbf{L}(t)^2 \rangle} \tag{3.14}$$

where brackets indicate ensemble averages over multiple realisations of walks and time t. It characterises the speed of displacement from the origin with time i.e., the diffusive nature of human mobility. For a two-dimensional random walk, we have $R(t) \sim t^{\frac{1}{2}}$.

There are a few classes of random walks: Brownian motion, Lévy flight, and Continuous time random walk. Empirical findings suggest that human trajectories are best described as Continuous time random walk (CTRW) [1]. CTRW is a random walk in which the number of jumps made in a time interval $\mathrm{d}t$ is also a random variable or equivalently, the time elapsed between jumps (Δt) is also a random variable which has a probability distribution of $\phi(\Delta t)$. And the the joint probability distribution function is $P(\Delta L, \Delta t) = f(\Delta L)\phi(\Delta t)$ due to the independence between Δt and ΔL.

Empirical results have suggested human trajectories have the below fat-tailed probability distribution of the jump length ΔL (trip distance) and the time difference between the origin and the destination Δt as illustrated in Section 3.2.2:

$$f(\Delta L) \sim \frac{1}{\Delta L^{1+\alpha}} \tag{3.15}$$

$$\phi(\Delta t) \sim \frac{1}{\Delta L^{1+\beta}} \tag{3.16}$$

where $0 < \alpha \leq 2$ and $0 < \beta \leq 1$. They are called Ambivalent Processes in CTRW which has $R(t) \sim t^{\frac{\beta}{\alpha}}$.

The nature of the diffusive behaviour is fully specified by α and β: for $\alpha < 2\beta$, the CTRW is super-diffusive and for $\alpha > 2\beta$, it is sub-diffusive; if $\alpha = 2\beta$ the random walk converges to ordinary diffusion/Brownian motion, despite the diverging moments of the respective distributions. [15].

If one side of human mobility is the diffusive nature, the other side of the coin is the returning effect i.e., people tend to return to one or more locations from day to day (preferential return). Song et al. [1] reveals the scaling properties of the number of distinct locations $S(t)$ as a function of time t follows $S(t) \sim t^{\mu}$ where $\mu = \beta$ for CTRW while they found $\mu < 1$. The rank-frequency of visited distinct locations follows a Zipf's law: $f_k \sim k^{-\zeta}$ where k is the rank of location according to the frequency of its being visited.

By combining these two sides of mobility, Song et al. [1] extended the CTRW model with the exploration and preferential return as briefly illustrated in Figure 3.7. They found:

$$\langle \Delta L^2 \rangle^{\alpha/2} \sim \log\left(\frac{1-S^{1-\zeta}}{\zeta-1}\right) + \text{const} \tag{3.17}$$

which relates the diffusion characteristic (MSD), $\langle \Delta L^2 \rangle^{\alpha/2}$, to the number of distinct locations S visited by an individual. This new model approximates the empirical data better than the other CTRW models. In Paper III, this mechanism of exploration and preferential return [1] serves as the core of the proposed model that fills the gaps in the sparse individual mobility traces. However, our model designs the details of the mechanism to accommodate the sparsity issue of the input data.

Another stream of individual mobility models stems from Transportation and Computer Science. These models further incorporate built environment, transport mode, and other social aspects of mobility using more sophisticated methods.

In the field of transport, activity-based models constitute a big category of travel demand models. Travel is the means to the end, that is participating in various activities. Given spatial, temporal and resources constraints, activity-based models predict the individual's activity chain in a certain time period that covers the number, sequence, and type of the activities [107], as illustrated by the space-time prism in time geography in Figure 3.8. In agent-based transport models, each agent's individual travel and the corresponding time-dynamic traffic is simulated at the microscopic level based on the transportation network and its attributes as the system constraints, where MATSim is a widely applied platform [108].

With the purpose of predicting individuals' whereabouts, some individual models are devoted to solving the problem of the next location prediction. This direction has a large number of applications, especially in context-aware services. For example, Do et al. [39] applied a probabilistic kernel

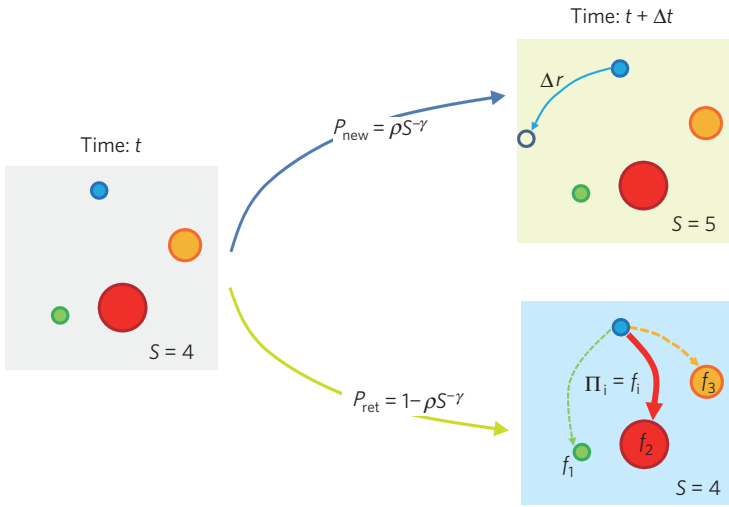


Figure 3.7: Schematic description of the individual-mobility model. Time t panel shows the starting time when historically an individual visited four locations, $S = 4$. Circles' size are proportional to their visiting frequency, f_i. For time $t + \Delta t$, this individual either visits a new location at distance Δr that follows a fat-tailed $P(\Delta r)$, or he/she returns to a previously visited location with probability $P_{\text{ret}} = 1 - \rho S^{-\gamma}$ where the next location will be chosen with probability $\Pi_i = f_i$. Source: Figure 2 from [1].

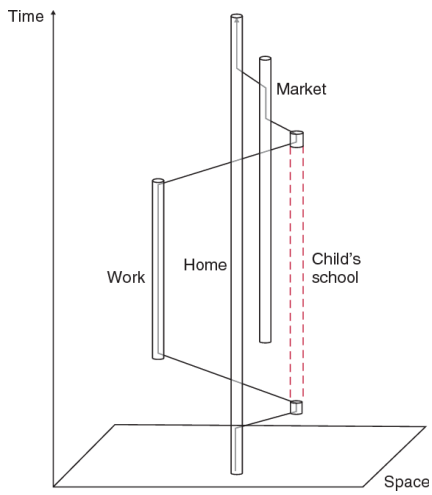


Figure 3.8: A space–time path among activity stations. Source: Figure 1 from [109].

method for human mobility prediction with smartphones. Other common methods include Markov models [110], dynamic Bayesian network, multilayer perception, and state predictor [111].

To summarise, this section briefly introduces the individual-level models that originate from a variety of disciplines including Complex Systems, Computer Science, and Transportation where the perspective of Complex Systems is more presented than the other perspectives due to the applied methods in the appended papers. More comprehensive reviews can be found in [15] on mobility physics, [90] on mobility models and machine learning, and [107] on big data and transport modelling.

3.2.4 Population-level models

The flows of the population between locations formulate an OD matrix that is modelled at the population level. This matrix has all possible combinations of origins and destinations for trips and it is easily transformed into a directed weighted network (**G**) in which nodes denote locations (for example counties or municipalities) and link weights correspond to the flow of travellers between the two locations [15]. The understanding of the mobility at the population level contributes greatly to Transport Geography and Urban Planning.

The Four-step model (FSM) is the primary tool for forecasting future demand and performance of a transportation system [112] as shown in Figure 3.9. Trip generation is the first step which estimates the number of trips produced by and attracted to each zone, either using empirical data directly or modelled results using zonal demographic and land use information. The step of trip distribution assigns trips produced by each zone to each of the other zones that these trips are attracted to [107]. After the first two steps, a total OD matrix is produced representing the population travel demand. Further through mode split and route choice, traffic flows are produced involving the transport system and traffic flow theories. The first steps are for population mobility modelling while the last two steps are in the scope of traffic flows modelling. This thesis focuses on the former aspect.

As the intermediate result of the first two steps in FSM, the total OD matrix estimates the number of trips $F_{i,j}$ from location i to location j from the socioeconomic characteristics of the populations of i and j, and their spatial distribution. Barbosa et al. [15] summarise a few mobility models to describe the total OD matrix. Distance-based models assume that the number of trips between two locations is a decreasing function of their distance, e.g., **gravity models**. **Intervening opportunities models** assume the number of potential destinations between two locations determines the mobility flow between them. **The radiation model** assumes the choice of a traveller's destination consists of two steps of "fitness evaluation".

The gravity model was first proposed in the 1940s to calculate mobility

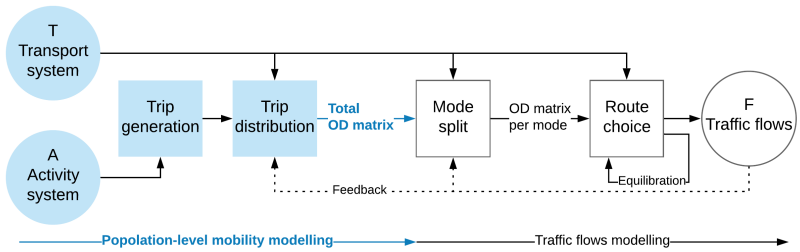


Figure 3.9: The Four Step Model. Adapted from Figure 2 in [112] and Figure 1 in [113].

flows inspired by Newton's law of gravitation [114] and later on became one of the most applied methods for the trip distribution [115]. The original form of the gravity model highlights the magnitude of $F_{i,j}$, a migratory flow between two communities i and j, has $F_{i,j} \propto \frac{P_i P_j}{r_{i,j}}$ where P_i and P_j represent the communities' population and $r_{i,j}$ the distance between i and j. A generic form of the gravity model is

$$F_{i,j} = k f_i f_j f\left(d_{i,j}\right) \tag{3.18}$$

where k is a constant, f_i and f_j are the number of produced trips (productions) and attracted trips (attractions) from zone i and to zone j respectively, and $f\left(d_{i,j}\right)$ the friction factor for travelling between zone i and j. There are many forms of the friction factor, one example used in Paper B is

$$f\left(d_{i,j}\right) = \alpha e^{-\beta d_{i,j}} \tag{3.19}$$

where $d_{i,j}$ can be the Haversine distance between the centroid of zone i and zone j or the other type of distance/travel time measures. In the real-world practice, getting the final total OD matrix also requires assigning trips from the predefined productions and attractions to each zone either as the origin or the destination. One example is called Iterative Proportional Fitting (IPF) [116, 117]. The parameters α and β are estimated or calibrated against some external data sources to minimise a certain form of error function between the model's estimates and the observed data.

Despite widespread use of the gravity model, it has notable limitations such as over-simplification and being data-demanding. Therefore, developing new models for the population mobility is a continuous effort. Intervening opportunities models proposed by Stouffer [118] have the main idea: "The probability that a trip ends in a given location is equal to the probability that this location offers an acceptable opportunity times the probability that an acceptable opportunity in another location closer to the origin of the trips has not been chosen." Along this track, the radiation model was proposed by

Simini et al. [119] and has been gaining increased attention. The job selection of the individual consists of two steps; 1) he/she seeks job offers from all counties (in the US) including his/her home county, and 2) the individual chooses the closest job to his/her home, whose benefits z are higher than the best offer available in his/her home county. As a result, the average flux $F_{i,j}$ from i to j predicted by the radiation model is

$$\langle F_{i,j} \rangle = f_i \frac{P_i P_j}{\big(P_i + s_{i,j}\big)\big(P_i + P_j + s_{i,j}\big)} \tag{3.20}$$

where P_i and P_j are the population in i and j and $s_{i,j}$ the total population in the circle of radius $r_{i,j}$ centred at i (excluding the source and destination population). Here f_i is the total number of commuters that start their journey from location i. This model is parameter-free and is particularly useful when there is a lack of previous mobility measurements and it significantly improves the predictive accuracy of most of the phenomena affected by mobility and transport processes.

Oftentimes we need to **compare two OD matrices** from different data sources or using different methods, especially when we want to know the validity of the emerging data sources or when we compare different models of population-level mobility. There are many ways to do this comparison. One newly proposed indicator is called **Spatially weighted structural similarity index (SpSSIM)** [120] as used in Paper B. SpSSIM is an extended version of the original structural similarity (SSIM) proposed by [121]. The original indicator was proposed to measure the similarity between two images for assessing image quality. This indicator was later introduced into the transport area for comparing the quality of OD matrices between data sources [122, 123]. This newly proposed SpSSIM [120] overcomes the SSIM sensitivity issue due to the ordering of OD pairs, as raised by previous studies [e.g., 124]. SpSSIM has a value between 0 and 1. SpSSIM equals 1 when two OD matrices have the exact same pattern.

The models mentioned so far aim at reproducing the observed mobility patterns at the population level. There are also some descriptive models designed for better characterising the patterns of population flows that are not easily observed from the raw OD matrix.

One descriptive model is the **community structure** which treats the OD matrix as a spatial network. In network science, a community is a group of nodes that have a higher likelihood of connecting to each other than nodes from other communities [101, p. 322]. In other words, a community is a locally dense connected subgraph in a network. Inspired by the question raised by Ratti et al. [2], "Do regional boundaries defined by governments respect the more natural ways that people interact across space?", the revealed community structure in human mobility has many applications, such as better placement and provisioning of services [125].

Using CDR datasets, the community structure detected displays a clear dis-

crepancy between the administrative boundary and the naturally formulated mobility partitioning (community structure). Huang et al. [126] compare different community detection algorithms in transport networks and find the Combo algorithm [127] outperforms the other algorithms, such as the Walktrap.

Therefore, in Paper V, in order to better understand the spatial patterns of ride-sourcing trips, the Combo algorithm is applied to the ride-sourcing ODM to detect the community structure [127]. This algorithm iterates over a sequence of moves that alter the community structure of the network to maximise the modularity gain and it can automatically decide the optimal number of communities [127]. As the result, the zones within a given community have a higher likelihood of connecting to each other by ride-sourcing than to zones in other communities (Figure 3.10).

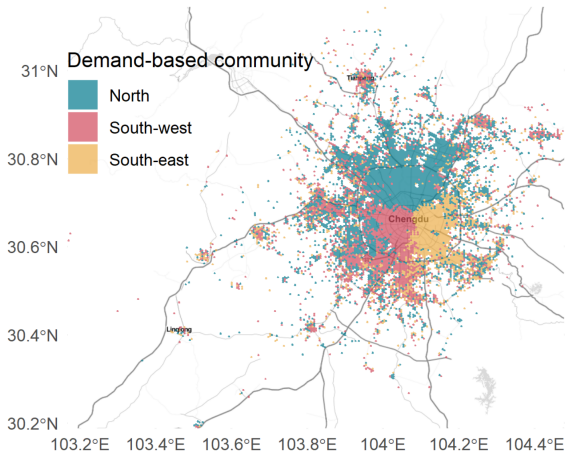


Figure 3.10: Spatial distribution of detected communities based on the ride-sourcing trips. Source: Figure 3(A) in Paper V.

To summarise, this section introduces the models of population-level mobility with the purpose of reproducing the OD matrix and the descriptive models taking community structure as an example. These models look into human mobility at the aggregate level producing significant insights of real-world relevance such as traffic modelling and urban planning.

3.3 Methods in transport geography

Transportation is interdisciplinary by nature. The methods in transport geography feature a reliance on empirical data and the intensive use of quantitative analysis ranging from descriptive measures to complex models [p304,

22]. Here, we introduce the main relevant methods used in the appended papers, while a more comprehensive introduction is presented in [22].

3.3.1 Routing with massive requests

Network analysis is one of the core methods used in this thesis, particularly routing by car and PT. In this part, **the shortest path problem** is a key function which is particularly useful for calculating travel time ($TT_{i,j}$). For example, we would like to know the modal disparities between car and PT on the answers to the question,"**How long does it take for one to go from anywhere to anywhere in Stockholm considering the real traffic at a given time?**" A request of routing contains an origin, a destination, and a departure time. Oftentimes, billions of such shortest-path calculation requests are required to be done efficiently. The routing by car and PT using open sources are introduced below, as they are the core methods in Paper IV. Using commercial APIs, an alternative when there is limited data access, is also introduced as used in Paper V.

Open-source solutions

For routing by car, we download drive road network from OSM and convert it into an igraph object [128] with edited links. Each link has the hourly average speed assigned as the routing impedance based on the speed records in the HERE Traffic data [74]. The calculation is implemented using python-igraph [128] where the Bellman-Ford algorithm is used to find the shortest paths between origins and destinations.

The complexity of the shortest path problem increases as we move from calculating travel time by car to PT, because it requires inter-modal routing to solve it. PT consists of many modes, e.g., walking, subway, and bus. To find the shortest travel time between two given locations by taking PT, the searching process must be done based on the multiple networks that are interconnected as well. GTFS data, as introduced in Section 2.3.2, are applied to calculate the travel time by PT.

OpenTripPlanner (OTP) is an open-source multi-modal routing engine [129], among various GIS solutions that support the routing process of taking PT [130–132]. A trip by PT potentially consists of all available modes of public transportation (bus, tram, train, subway, etc.) and walking. For each pair of origin-destination, OTP finds the fastest door-to-door trip given a set departure time and the combination of transport modes available. Many parameters e.g., the maximum walking distance and the walking speed, are configurable. In Paper IV, the downloaded GTFS data and map data from OSM are fed into OTP API via a Python script where we implement the parallelisation of billions of OD requests.

When calculating the travel time by PT, the modifiable temporal unit prob-

lem (MTUP) needs careful attention. MTUP is defined as the effects of temporal aggregation, segmentation, and boundary. When creating the requests that contain origin, destination, and departure time, the selection of departure times will affect the results of aggregate travel time. These are the result of the interaction between two dimensions: the temporal sampling strategy and the temporal sampling frequency [133]. Therefore, in Paper IV, to better balance the errors and the computation time, we use a hybrid sampling approach with 15-min resolution [133].

Using commercial APIs

The above methods rely on data such as road networks and GTFS data. When these data are not available, one may switch to commercial APIs for their map services. For instance, Google Map Direction API allows free access of up to 2,500 requests per day. One can get live traffic and travel time estimation for travelling by car or PT. In Paper V, in order to get the travel details of taking PT, we feed the pick-up time and the pick-up and drop-off locations of each ride-sourcing record into Baidu Transit API [134]. The API returns the trip information for taking PT including travel time, total walking distance, and the number of boardings etc.

3.3.2 Hotspot analysis

Hotspot analysis is a spatial analysis to identify clustering of spatial occurrences as points in a map, such as crime, traffic crashes, and pick-up and drop-off locations of trips. Unlike the density map describing the data, in such analysis, the statistically significant hot/cold spots will be identified distinguishing them from the rest of the study area. A hot spot is an area that has a greater than the average number of event occurrences. Figure 3.11 shows an example of hotspot analysis of the pick-ups of the ride-sourcing trips in Paper V.

Getis-Ord Gi* is a method of detecting hotspots by looking at the zones in the dataset within the context of the neighbouring zones in terms of the event occurrence [135]. As a result of the analysis, a z-score and a p-value are returned for each zone in the study area. The statistically significant spots at 90% confidence level are those zones with $p < 0.1$ and z-score > 1.65 for hotspots while coldspots have z-score < -1.65. Similarly, for 95% confidence level, the corresponding p-value below 0.05 and z-scores have either < -1.96 or > 1.96. Getis-Ord Gi* is applied in Paper V to detect the hotspots of pick-up and drop-off locations of the ride-sourcing trips. As illustrated in Figure 3.11, not all the zones with great number of pick-up count can be called statistically significant hotspots (90% confidence level).

Spatial zones of doing this analysis are critical. Similarly to MTUP, we have the modifiable areal unit problem (MAUP) at the spatial dimension [132].

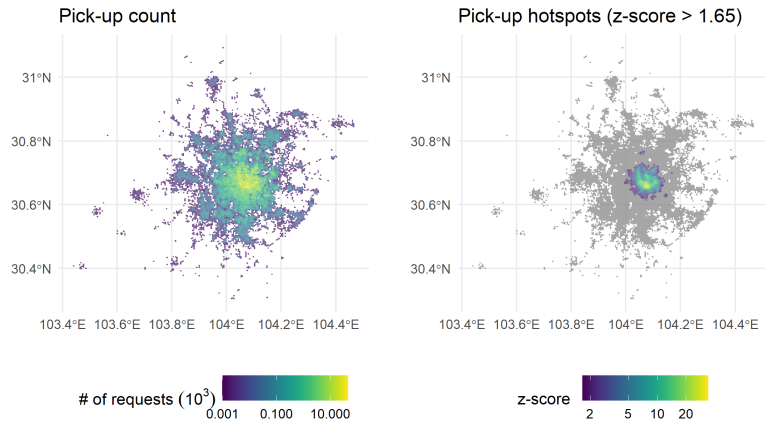


Figure 3.11: Spatial distribution of pick-up counts (right) and pick-up hotspots (left, coloured) based on the ride-sourcing trips. Source: study of Paper V.

Depending on how the study area is divided, the detected hotspots tend to be different. Therefore, as opposed to zip-code zones, if data availability allows, it is recommended to create UTM-based grid zones that have a regular shape and uniform area, such as the hexagonal zones in Paper IV and V.

CHAPTER 4

Modelling potentials and limitations of mobility data

Using emerging data sources, particularly Twitter data, Paper I-III demonstrate the process of understanding mobility and further apply the obtained knowledge of mobility in the field of transport. They attempt to answer the below question:

- What are the potentials and limitations of using these emerging data sources for modelling mobility?

An overview of the research scope and the involved data sources are presented in Figure 4.1. Paper I [10], Paper II [11], and Paper III [12] reveal the potentials and limitations of using geolocations of Twitter data for modelling mobility.

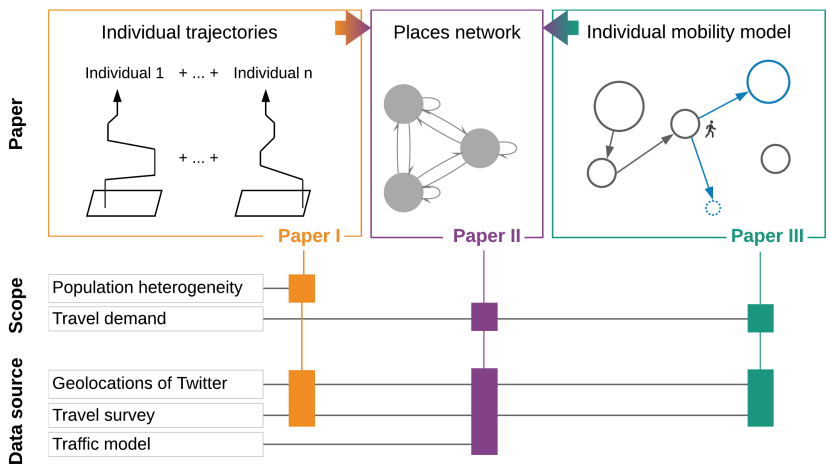


Figure 4.1: Overview of included studies: their scope and involved data sources.

Paper I focuses on the aspect of individual trajectories to reveal if the population heterogeneity on the spatiotemporal patterns of mobility can be

captured by geolocations of Twitter. Paper II focuses on the travel demand estimation (places network) aggregating individual trajectories, particularly the feasibility of using geolocations of social media data for travel demand estimation. Paper II examines the effects of data sparsity, spatial scale, sampling methods, and sample size on this feasibility. Paper III extends the use of these low-cost and easy-to-access emerging data, by proposing a model to fill the gaps in sparse individual traces for travel demand estimation. All three papers validate the results from Twitter data against some established data sources.

The following sections provide a summary of the appended papers on their motivations, research questions and methods, main findings, and conclusions.

4.1 Population heterogeneity of mobility (Paper I)

From individual to collective behaviours: exploring population heterogeneity of human mobility based on social media data

Motivation

Literature review suggests a two-fold research gap in the use of Twitter data. First, most studies use lateral geotagged tweets that are collected from Streaming API (more details in Section 2.3.1) and therefore, focus on the mobility that happens within a small area while the movements across the geographic boundary are not captured. Second, most studies of aggregate population behaviours neglect individual differences, while studies of individual mobility usually neglect common features that drive similar behaviours across groups of individuals; there has been little work on combining aggregate and individual perspectives to gain new insights about travel behaviours of a heterogeneous population. And this heterogeneity sheds light on a more sophisticated mobility modelling in many disciplines such as epidemics and urban planning. However, the feasibility of using geotagged tweets to represent the population heterogeneity remains unclear.

Research questions and method

This paper reveals the population heterogeneity of geotagged activity patterns using a long-term dataset without any geographical boundaries, such as national borders or administrative boundaries. Specifically, this study attempts to answer the following three questions.

- Are there any distinct patterns that characterise the observed individual geotagged activities?

- What are the spatial and temporal characteristics derived from different geotagged activity patterns?
- Can geotagged tweets be used as a proxy to approximate the mobility patterns of different behavioural groups?

To answer these questions, we use three datasets. Twitter dataset, from User Timeline API (more details in Section 2.3.1), includes more than 650,000 geotagged tweets by nearly three thousand Swedish Twitter users covering time spans of more than 3 years on average. For the sake of validation, we also collect individual trip information from the Swedish National Travel Survey and the population distribution from the up-to-date census data in Sweden. We use the travel survey data to investigate the representativeness of geotagged tweets via a descriptive analysis, comparing spatio-temporal characteristics (behaviour distortion) and the population distribution (population biases).

To identify the population heterogeneity of geotagged activity patterns, we combine aggregate and individual analysis techniques: we first analyse the geotagged trajectories of each user to classify them regarding their activity patterns, and then we conduct an aggregate analysis for each group. We characterise the features of individual trajectories of geotagged tweets using both geographical and network properties. The features describing users' activity patterns are based on those found in the literature. Hierarchical clustering, a descriptive data mining method is used to produce new, non-trivial classifications of users based on their set of features.

Main findings

Validation: Twitter vs. survey and census

As introduced in Section 2.3.1, behaviour distortions and population biases are two main disadvantages of Twitter data. To fully acknowledge the limitations of the geotagged tweets, we first show the differences in the descriptive characteristics between Twitter data and the other two data sources, the travel survey and the census data in Figure 4.2.

One significant observation is about the population biases (Figure 4.2A-B). **Compared with the general population, the top Twitter users in Sweden seem to over-represent the residents in big cities**, especially the capital city in Stockholm county, while the rest of the top Twitter users seem to be distributed similarly to the population distribution and the participants in the travel diary.

Another aspect of the findings is the behavioural distortion (Figure 4.2C-E). The ratio of distinct locations quantifies the variation level of geotagged locations for each user (Figure 4.2C). The more geotagged locations that are outside the habitually visited locations, the larger the variation level.

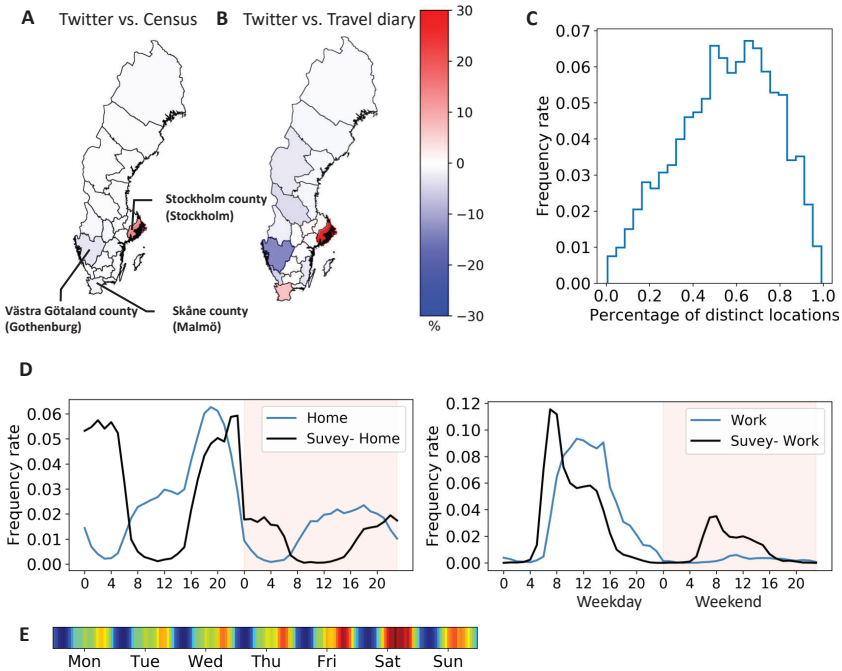


Figure 4.2: Characteristics of geotagged activity of Swedish Twitter users (adapted from Figure 2 and Figure 3 in Paper I). (A) and (B) show the county-level geographical representativeness of estimated home locations from Twitter data: percentage value difference. (**A**) Twitter users vs. residents (Twitter minus Census population). (**B**) Twitter users vs. Swedish travel survey participants (Twitter minus survey). (**C**) The distribution of the ratio of distinct geotagged locations over total geotagged locations (individually calculated). (**D**) Daily distributions of visiting frequency of the top two most visited locations, weekday vs. weekend (adjusted by the overall distribution of geotagged tweeting frequency over seven days across a week). (**E**) A week-long geotagging activity pattern (average of all the users). The warmer the colour (e.g. red and orange), the higher number of geotagged locations.

We further assume that the first and the second most visited locations by users are either work or home. These two locations have distinct temporal distributions in a day. We apply a hierarchical clustering to the instances of users' daily time distribution of visiting frequency for these two locations. We find two significantly different patterns that fit work and home respectively (Figure 4.2D). At the same time, we also observe that geotagged tweets tend to represent the activities that happen during lunch time and night (Figure 4.2E).

If users constantly and regularly tweet during a certain daily time frame or only from a few selected locations, then the locations we capture are skewed to the locations that they tend to visit during that time frame. How-

ever, as seen in our study (Figure 4.2C), **it is not the case that people only geotweet from a few fixed locations**. Despite peaks during lunch time and night (Figure 4.2E), **geotagged tweets capture many routine activities** (Fig. Figure 4.2D), as seen from the temporal profile of the first and second most visited locations that share some similarities with the "ground truth" in the travel survey.

Four distinct groups of travellers: population heterogeneity on mobility

After the descriptive analysis of comparing Twitter data with the travel survey and census data, we identify **four distinct behavioural groups of Twitter users on their mobility patterns** as summarised in Figure 4.3. The six features are defined to describe the individual trajectory of geotagged tweets. Among them, geographical characteristics refer to the travel distance range (weighted by the visiting frequency), location distance variance, and the average distance between two consecutive geotagged tweets. And network properties are to which degree the visited locations are connected together, the overall visiting frequency, and the degree of how centralised the overall visited locations are from visiting frequency. In short, **mobility is described in two aspects: how far one travels and how frequently one explores new locations**.

The statistical summary of the four behavioural groups is shown in Table 4.1. It shows an imbalanced distribution of Twitter users across four groups. **Most users are local returners who mostly geotag locations that are within Sweden**. A high returning rate and frequent geotweeting behaviour are associated with the centralised network structure of geotagged locations which distinguishes returners and explorers. However, the later test has ruled out the effect of geotweeting frequency on the clustering results. In other words, **the identified four groups are not sensitive to the change of geotweeting frequency**.

For the collective mobility behaviours, we further show their trip distance distribution and how different groups diffuse in space in Figure 4.4.

The trip distance generally increases with the waiting time over a multiple-day period at a decreasing rate to up to 7 days (Figure 4.4A-B). **The diffusive nature of human mobility and the returning effect (e.g., return to home or return to work) create two distinct mechanisms that interact with each other**: the diffusion effect causes the observed trip distance to increase with increasing waiting time derived, and the returning effect causes some of the distances to decrease to zero periodically, i.e., every 24 hours. **Diffusive effect sustains longer in explorers compared with returners because they are more active on exploring new locations**.

The cumulative frequency rate reflects the regularity of users' visiting behaviour. **Returners have more concentrated visits to a fewer number of**

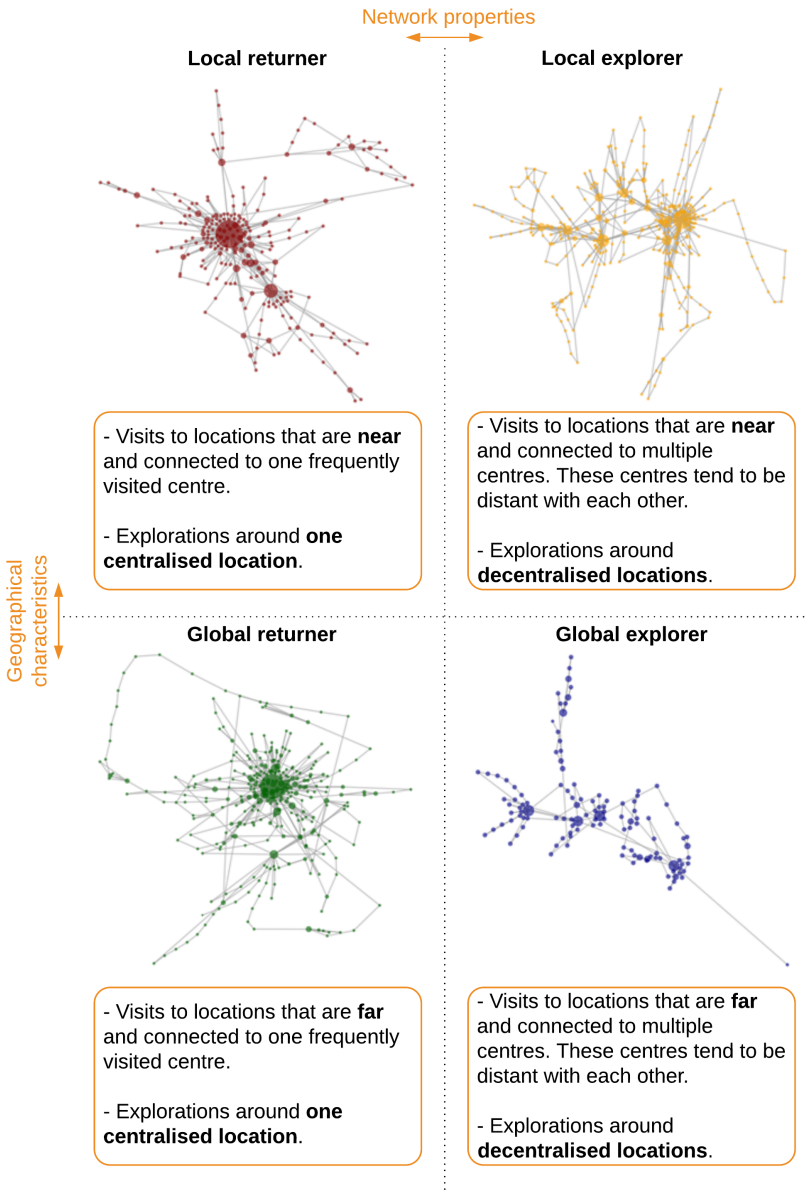


Figure 4.3: Network visualisation of four representative individuals from each behavioural group and a brief summary of the group characteristics (adapted from Figure 5 in Paper I). In the visualised networks, each node represents one visited location. The diameter of the node is proportional to the node degree.

Table 4.1: Statistics of four behaviour groups. dom represents the percentage of trips where both the origin and destination are in Sweden (0), among the destination and the origin, there is one location outside Sweden (1), and both the origin and destination are outside of Sweden (2). R denotes the ratio of visiting frequency of the most frequently visited location over the total number of geotagged locations. F_g denotes the geotweeting frequency.

Name	User (%)	dom (%)			R	F_g (/day)
		0	1	2		
Local returner	14.4	81.3	7.0	11.7	0.4	0.6
Local explorer	78.0	88.4	5.0	6.6	0.2	0.3
Global returner	0.3	45.9	10.0	44.1	0.4	1.6
Global explorer	7.3	39.6	12.1	48.3	0.2	0.3

locations than the explorers do (Figure 4.4C). According to the diffusion process in space, **the global travellers have a larger mobility range than the local travellers which increases continuously throughout the time period, whereas the local travellers' mobility range tends to saturate earlier** (Figure 4.4D).

Conclusions

Paper I explores the population heterogeneity of spatial mobility including travel and day-to-day displacement, from a combined perspective of individual actors and collective behaviours. The findings of this paper could be relevant for disease prediction, transport modelling, and the broader social sciences.

Our analysis framework provides a coherent picture of the geotagged activity patterns by combining the individual perspective with the aggregate perspective. We use a social media dataset of 652,945 geotagged tweets generated by 2,933 Swedish Twitter users covering an average time span of 3.6 years. No explicit geographical boundaries, such as national borders or administrative boundaries, are applied to the data. We use spatial features, such as geographical characteristics and network properties, and apply a clustering technique to reveal the heterogeneity of geotagged activity patterns. We find four distinct groups of travellers: local explorers (78.0%), local returners (14.4%), global explorers (7.3%), and global returners (0.3%). These groups exhibit distinct mobility characteristics, such as trip distance, diffusion process, percentage of domestic trips, visiting frequency of the most-visited locations, and total number of geotagged locations.

Geotagged social media data are gradually being incorporated into travel behaviour studies as user-contributed data sources. While such data have many advantages, including easy access and the flexibility to capture movements across multiple scales (individual, city, country, and globe), more attention is still needed on data validation and identifying potential biases

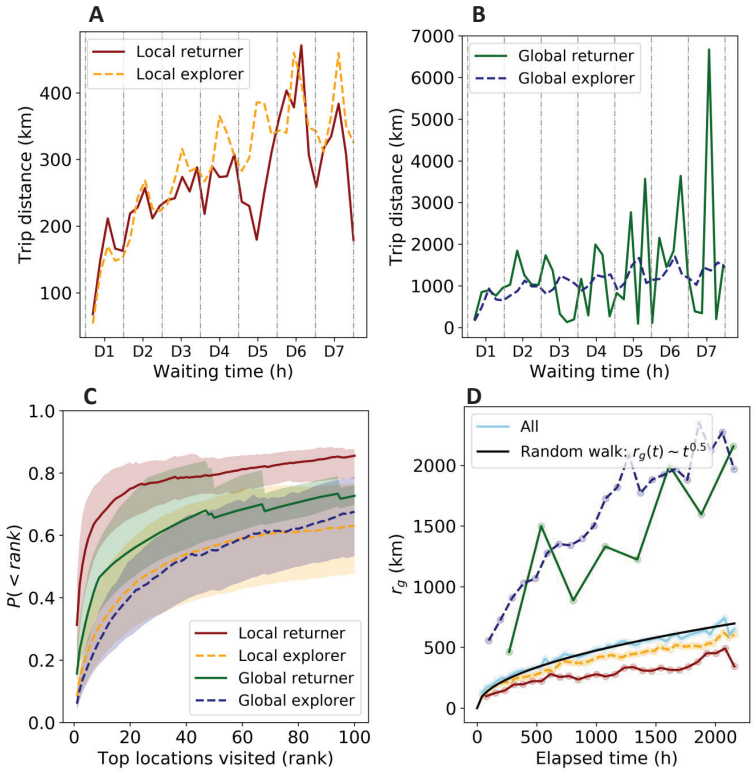


Figure 4.4: Collective mobility behaviours (adapted from Figure 9 and 10 in Paper I). Trip distance vs. waiting time during 7 days for (**A**) local travellers and (**B**) global travellers. Waiting time is defined as the time interval between two consecutive geotagged tweets generated by the same Twitter user. (**C**) Cumulative visiting frequency by the ranking order of the top 100 visited locations. The shaded range indicates the upper bound (75%) and lower bound (25%) of the cumulative frequency rate of visits. (**D**) Time history of radius of gyration within 90 days. The time history starts from the first time observing the most visited location; each data point indicates the mean value of radius of gyration across the same group of users.

associated with these data. We validate against the data from a national travel survey and find that despite good agreement of trip distances (one-day and long-distance trips), we also find some differences in home location and the frequency of international trips, possibly due to population bias and behaviour distortion in Twitter data. Future work includes identifying and removing additional biases so that results from geotagged activity patterns may be generalised to human mobility patterns.

4.2 Travel demand estimation (Paper II)

Feasibility of estimating travel demand using geolocations of social media data

Motivation

Travel demand estimation, as quantified by origin-destination (OD) matrix is essential for urban planning and management of transportation networks. In the last decade, emerging data sources have significantly improved our understanding of travel behaviour. Among them, the low cost makes geotagged tweets appealing for the travel demand estimation, especially when the traditional data sources, e.g., census and road surveys, are increasingly costly and hard to keep up-to-date. There is also a consensus on the need for careful inspection of using geotagged social media data to approximate the travel demand patterns from established data sources.

The work comparing geotagged tweets with other data sources for travel demand estimation still lacks systematic rigour in at least four areas: 1) **Commuting travel demand**. The basic temporal technique to identify home or workplace has been widely applied for deriving commuting trips. Our preliminary results from previous analyses suggest that identifying home and workplace locations through geotagged tweets gives mixed results and the reliability of the method requires further scrutiny; 2) **Spatial scale**. Most studies look at pre-selected regions without exploring the effects of spatial scales on travel demand estimation, whereas we hypothesise that the feasibility of using Twitter data for travel demand estimation can depend on the scale; 3) **Sampling methods**. The existing literature is not clear on how different sampling methods (**region-based, Twitter LT vs. user-based, Twitter LD**) affect the validity of using geotagged tweets to estimate travel demand; 4) **Sample size**. It remains unclear how the sparsity of Twitter data affects the validity of using it for travel demand estimation.

Research questions and method

Paper II comprehensively examines the validity of using geotagged tweets collected within a specified region, and from user timelines, to approximate the OD matrix at different spatial scales. We compare these Twitter-based OD matrices with the Swedish national travel survey and output from Swedish Transport Administration (Trafikverket) traffic models. Specifically, we attempt to answer the following questions:

- Are Twitter data a feasible source for representing commuting travel demand?
- Can geolocations of Twitter data be used to create models for travel

demand estimation?

- How do spatial scale, sampling method, and sample size of Twitter data affect its representativeness for travel demand?

We first compare the empirical trip records with respect to the commuting travel demand and the overall travel demand for an average weekday. We then create gravity models based on Twitter data to estimate the overall travel demand at both the national (long-distance travel above 100 km) and city level. Finally, we compare Twitter-based OD matrices and trip distance distributions with those from the other established sources using spatially weighted structural similarity index (SpSSIM) and Kullback-Leibler divergence (KL divergence), respectively.

Main findings

Commuting travel demand estimation

The reliability of estimated commuting trips using geotagged tweets is low.

As shown in Figure 4.5, the commuting OD using Twitter data and the one based on Survey are **not similar** according to the visual result, the similarity metric (SpSSIM = 0.39), and the commuting distance distribution (KL divergence = 0.052).

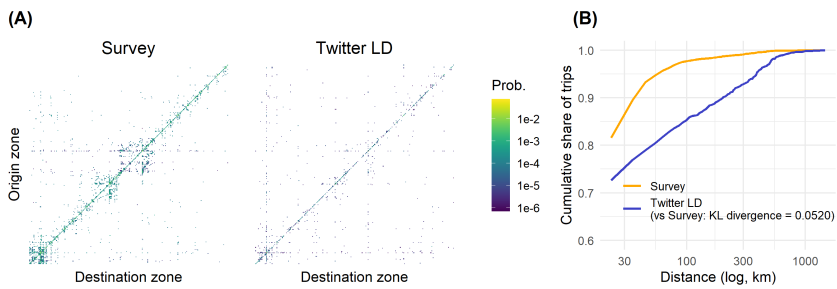


Figure 4.5: Evaluation of the feasibility of using Twitter for commuting travel demand estimation (adapted from Figure 4 and 5 in Paper II). Commuting OD matrices based on (**A**) Survey and user-based collected geolocations (Twitter LD). (**B**) Commuting trip distance distribution produced by Twitter LD in comparison with Survey.

Twitter data itself does not include any location information. Therefore, it is common to use the temporal profiles of being at home and workplace to identify these locations that are potentially included in the individual trajectory of geotagged activities. However, the estimated home and workplace based on Twitter LD are not reliable. One explanation is that most Twitter users may not feel comfortable to post their home and workplace online

publicly due to privacy concerns. Twitter users' temporal distribution of geotagging behaviour resembles a leisure activity pattern as also confirmed in Paper I. Moreover, geotag users tend to geotag locations that are not within their neighborhood; and the geotagged locations concentrate substantially at locations farther away than the daily mobility area. These evidence point to the fact that Twitter data has a low representation of routine activities such as visiting the workplace or home.

The impact of spatial scale

The main obstacle of using Twitter data at a large spatial scale is the sparsity.

At the national and city level, the similarity between the OD matrices based on Twitter data and the Sampers' model output (as ground truth) is shown in Table 4.2. Paper II illustrates the model outcomes are visualised in Figure 4.6 and the distance distribution of the model outputs in Figure 4.7.

Table 4.2: The similarity between the modelled OD matrices using Twitter data and Sampers traffic model's outputs. Model A: displacement conversion plus gravity model; Model B: density-based approach plus gravity model. For all models, $\beta = 0.03$.

Scale	Model	Twitter	SpSSIM	KL divergence
City	A	LD	0.74	0.072
		LT	0.54	0.219
	B	LD	0.80	0.023
		LT	0.85	0.017
National	A	LD	0.52	0.317
		LT	0.40	0.364
	B	LD	0.54	0.021
		LT	0.54	0.026

Twitter data is more suitable for estimating the overall travel demand at the city level compared to the national level (long-distance travel) in terms of similarity and the distance distribution. Twitter data generally work well at the city level (0.54 to 0.85), while the performance at the national level is not as good (0.40 to 0.54), see Table 4.2. The sampling method matters; Twitter LD is more similar to Sampers than Twitter LT, especially when using Model A with Displacement conversion (National, Twitter LD^A vs. Twitter $LT^A = 0.52$ vs. 0.40 and City, 0.74 vs. 0.54). Combining the density-based approach and the gravity model (Model B) produces better similarity results compared to using Displacement conversion (Model A) at both spatial scales. This is probably due to the fact that Model B manages to increase the number

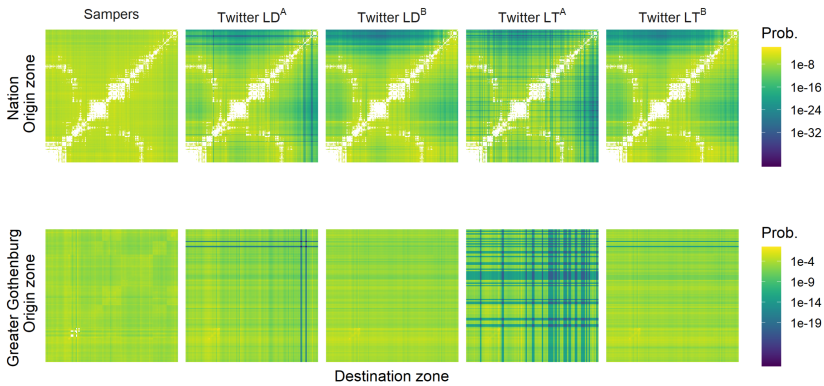


Figure 4.6: Estimated OD matrices by gravity model and Sampers' model outputs. ^{A}: Displacement conversion plus gravity model. ^{B}: Density-based approach plus gravity model. Source: Figure 8 in Paper II.

of available geotagged tweets five-fold relative to Model A. Therefore, using geotagged tweets for travel demand estimation requires reasonable spatial aggregation which depends on the form of Twitter data and the penetration of Twitter.

The impact of sampling size and methods of data collection

The more geotagged tweets included in the modelling, the better Twitter is at estimating travel demand. User-based data collection results in a much larger number of geotagged tweets that overall better represents population mobility patterns.

Sensitivity of outcomes to sample size and to sampling method of tweets (LD or LT) are tested using a share of geotagged tweets from 1% to 99%, with a step length of 1% and 10 repetitions of random sampling, to create outputs using models A and B with the same settings as above. Figure 4.8 shows the similarity results. As more geotagged tweets are included in the modelling, the similarity between the outputs of the Twitter-based OD matrix and the Sampers model increases and remains within a smaller range. The national level is more sensitive to data sparsity, because the number and the geographical coverage of traffic zones is greater than at the city level, therefore requires a greater number of tweets to reach a stable (but still lower) similarity. In terms of methodology, Model A is more sensitive to the number of geotagged tweets than Model B, especially with respect to the stability of the results with a smaller number of tweets, and is generally associated with poorer results.

Compared with Twitter LT, Twitter LD covers a longer period (9 years compared with 6 months for Twitter LT) with fewer users (2,311 compared with

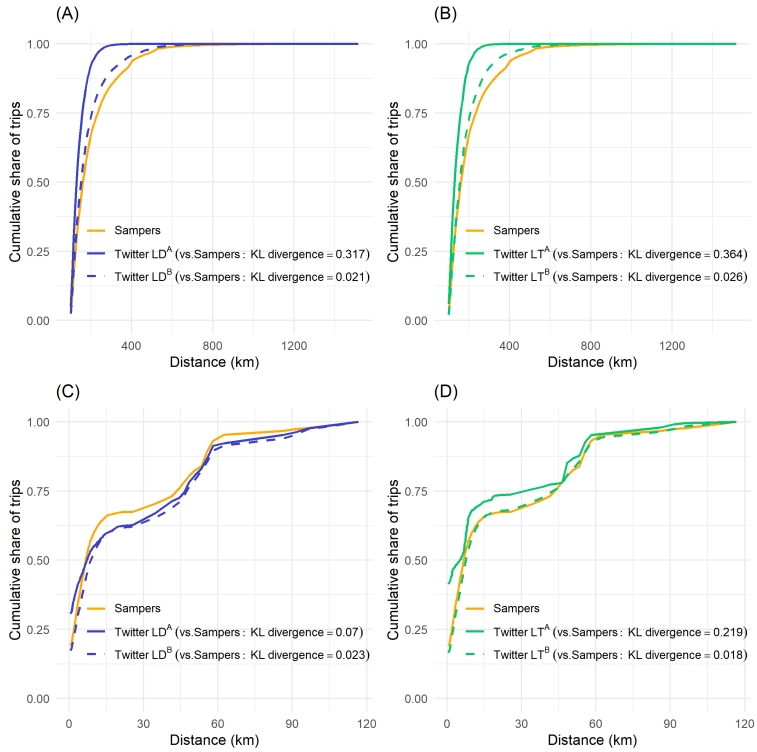


Figure 4.7: Trip distance distribution. Cumulative share of trips refers to the probability of travel between zones below a given distance. The trip distance is from the estimated OD matrices by *A* displacement conversion plus gravity model and by *B* density-based approach plus gravity model. (**A**) National level - Twitter LD. (**B**) National level - Twitter LT. (**C**) City level - Twitter LD. (**D**) City level - Twitter LT. Source: Figure 9 in Paper II.

24,442 with Twitter LT). Our study demonstrates that, however, the long-term coverage of longitudinal geotagged tweets by top users (User Timeline API) compensates for the time sparsity and helps to recreate a more complete picture of population mobility patterns, and therefore, is more reliable for travel demand estimation than the lateral dataset (Twitter LT). However, this gap narrows or disappears when using a novel density-based approach developed in this study.

A novel density-based approach: geotagged tweets as attractions generators as opposed to trips generators

The density-based approach utilises more geotagged tweets, resulting in better representation of travel demand.

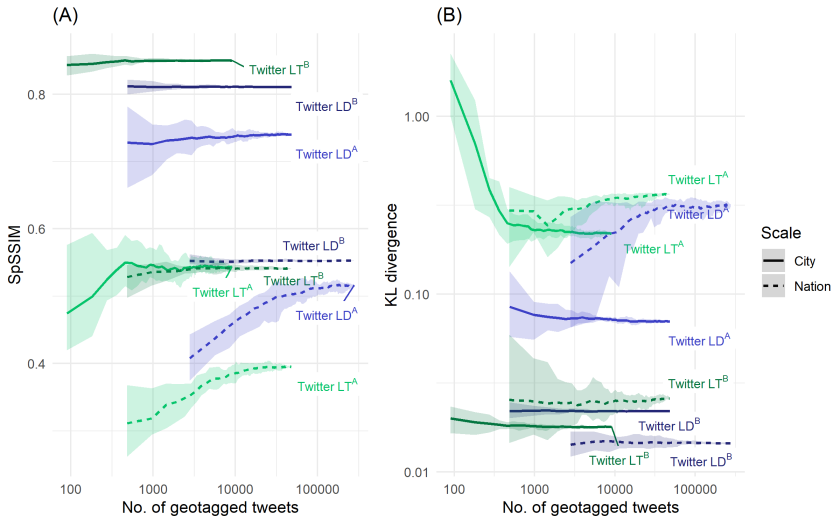


Figure 4.8: Similarity, (**A**) SpSSIM and (**B**) KL divergence, as a function of the number of geotagged tweets. Green colours show the results using Twitter LT and blue colours show the results using Twitter LD. For the 10 model runs of each tweets sample size, the curve shows the average value of SpSSIM/KL divergence and the shaded area shows the maximum and minimum value of SpSSIM/KL divergence. Model A - displacement conversion plus gravity model; Model B - density-based approach plus gravity model. For all models, $\beta = 0.03$.

Twitter users geotweet to report activities instead of trips, therefore, the density of geotagged tweets naturally reflects the attractiveness of zones. In the density-based approach (Model B), we assume the generated trips between zones are proportional to 1) the population and 2) the number of activities some of which are geotagged. The proposed density-based approach regards the tweets density of zones as the attractions and the population size of zones as the productions.

On the other hand, Model A - displacement method, a common practice of adding a time threshold to capture "trips", drastically reduces the available Twitter data for travel demand estimation: only 20-35% of geotagged tweets are utilised to estimate the overall travel demand. This reduction limits the application of geotagged tweets given that sparsity is already one of its key drawbacks.

As a comparison, without the need for a time threshold, the density-based approach (Model B) increases usable data by 2-7 times. This drastically increases the similarity scores of the OD matrices compared with Sampers' model outputs and the method is not so sensitive to sample size. Not only does the density-based approach produce better OD matrices, but it also produces better trip distance distributions compared with the Surveys.

The density-based approach can be extended to compute time-dependent

attractions by aggregating geotagged tweets across different temporal profiles, providing a dynamic picture of travel demand by time of day, week, or season.

Conclusions

This study systematically explores the feasibility of using geolocations of Twitter data for travel demand estimation by examining the effects of data sparsity, spatial scale, sampling methods, and sample size. We show that Twitter data are suitable for modelling the overall travel demand for an average weekday but not for commuting travel demand, due to the low reliability of identifying home and workplace. Collecting more detailed, long-term individual data from user timelines for a small number of individuals produces more accurate results than short-term data for a much larger population within a region. We developed a novel approach using geotagged tweets as attraction generators as opposed to the commonly adopted trip generators. This significantly increases usable data, resulting in better representation of travel demand.

The key strengths of social media data are that they are low-cost, abundant, available in real-time, and free of arbitrary geographical partition. However, there are also significant limitations: population and behavioural biases and lack of important information such as social demographic information and trip purposes. Despite clear indications of overly representing residents in big cities and their leisure activities from the existing literature, we demonstrate in the present study that geotagged tweets can provide a reasonably good travel demand estimation that also captures the trends over time, though careful consideration must be given to sampling method, estimation model, and sample size.

4.3 Synthetic travel demand by a mobility model (Paper III)

A Mobility Model for Synthetic Travel Demand from Sparse Individual Traces

Motivation

Transportation presents a major challenge to curbing climate change. Meeting the challenge will require knowing the details of travel demand, how and how much people travel. The mobility traces from these sources are important in quantifying the flows of people between places and how far they travel [15]. One salient issue is to what extent the covered traces are incomplete, i.e., the sparsity issue. The incompleteness limits the accuracy of the estimated travel demand. Given that geolocations are collected with

triggered phone activities or volunteered reports, data sources like CDRs, LBSNs, and social media data only provide a partial view of the actual mobility trajectories [103]. However, these sources are collectively abundant, especially LBSNs and social media data, which are also inexpensive and easy to access.

Most studies directly use them and ignore the impact of the sparsity issue, leading to results that are potentially biased and inaccurate. The efforts of filling the gaps in the sparse traces have mainly been applied to CDRs, while increasingly popular sources such as LBSNs and social media data are rarely considered. In order to extend the use of these inexpensive and easy-to-access data, it is crucial to design appropriate techniques to fill the gaps in sparse mobility traces. By doing so, one can deliver a more reliable synthetic travel demand.

Research questions and method

Paper III proposes an individual-based mobility model to deal with sparse mobility traces, particularly geolocations of social media data. We calibrate and validate the model with the other established data sources in the form of origin-destination matrices quantifying the population travel demand in Sweden, the Netherlands, and São Paulo, Brazil. We demonstrate the usefulness of the model in characterising domestic trip distances for 22 global regions including cities and countries. Specifically, we attempt to answer these nested research questions:

- How to develop a model that fills the gaps in sparse mobility data for a more accurate synthetic travel demand?
- How well does the model perform on the validation data, with parameters calibrated by region?
- Can the calibrated model be applied to new cities or countries?

The proposed model is shown in Figure 4.9. The model-synthesised data can simulate the population flows and characterise the trip distance distribution. The proposed model applies the mechanism of exploration and preferential return as its core when synthesising mobility traces [1]. However, our model designs the details of the mechanism to accommodate the sparse individual traces.

We use the visitation frequency obtained by using Zipf's law when designing the probability function for returning to an old place instead of the one directly calculated from the sparse input. This is because the long-term observation of individual geolocations of Twitter users captures both routine mobility and occasional exploration to new places [10], despite the proportion of regular locations to uncommon places deviating from the users' actual mobility [50, 65]. In doing so, we attempt to exclude the bias of overly rep-

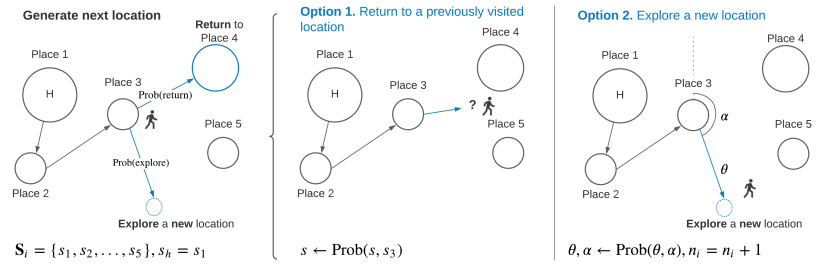


Figure 4.9: Model framework of generating mobility traces. An example of individual i with 5 distinct observed locations. He/she is currently located at Place 3 (s_3). Source: Figure 1 in Paper III.

resenting uncommon places in the sparse geolocations of Twitter users. We also create a two-dimensional distribution of jump size (trip distance) and bearing for exploring a new place, instead of replicating the biased displacements in the sparse traces. This distribution is shaped by the individual's returning and exploring behaviour observed in the Twitter data, and the visits to new places are constrained by where the individual lives and stays most of the time. With this new model, the sparse individual mobility traces are synthesised into a more representative set for synthetic travel demand.

An example individual from São Paulo is shown in Figure 4.10. In Figure 4.10(a), the sparse geolocations of Twitter data are mainly distributed in central São Paulo, however, the time intervals between any two consecutive geolocations reported by this Twitter user are much longer than the time they actually spend on travelling between them because the departure and arrival times are not precisely logged. However, the model fills the gaps in Figure 4.10(a) so that we can connect those visits to form synthetic trips that spread across the study area (Figure 4.10(b), small chart at the top-right corner). Nevertheless, most trips are located in the sub-area where the sparse traces concentrate, as shown in the main chart of Figure 4.10(b).

To test the model, we use geotagged tweets as an example of sparse traces. We first construct models for Sweden, the Netherlands, and São Paulo and calibrate the models against the official travel survey data as the "ground truth" to find the optimal parameters. The aim of the experiment is to see how the model performs in representing the travel demand as quantified by the population flows when validated against the other established data sources. The model performance is evaluated by comparing the ODM and its trip distance distribution with the ground truth. To illustrate the usefulness of the model, we apply the validated model to the sparse traces collected from 22 regions including countries and cities for which we have the geolocations of Twitter data to create synthesised domestic trip distance distributions.

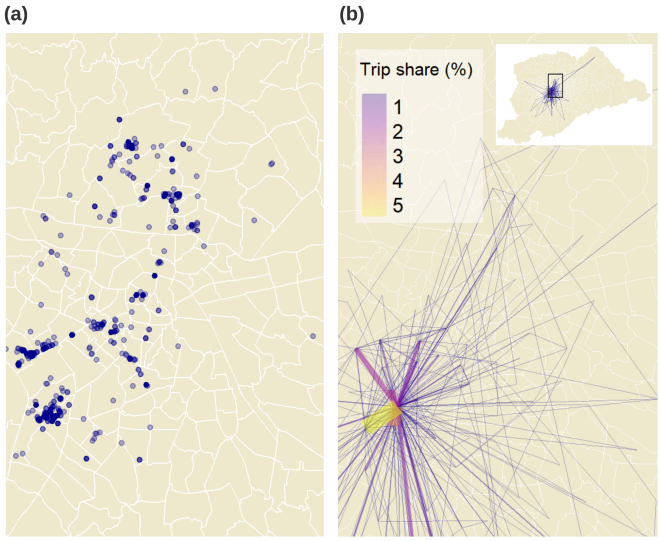


Figure 4.10: Model input and output: a selected individual from São Paulo, Brazil. (a) Sparse individual traces. (b) Model synthesised traces. The warmer the colour, the higher trip frequency between spatial zones. Source: Figure 5 in Paper III.

Main findings

Representing travel demand

The model produces synthetic origin-destination matrices and trip distance distributions that have good agreements with the other data sources on origin-destination matrices and trip distance distributions.

Aggregating the model output of all the Twitter users and the trips in the ground truth data, we quantify the population flows between the spatial zones in the study areas and the corresponding trip distance distributions in Figure 4.11.

The model generally performs better for OD pairs of higher frequency rate than for those of lower frequency rate. And the performance varies between the three regions. Taking the average correlation between the ground truth and the model output (validated and calibrated), the proposed model performs the best in São Paulo (Kendall's tau = 0.32, $p < 0.001$), followed by the Netherlands (Kendall's tau = 0.19, $p < 0.001$), and Sweden (Kendall's tau = 0.12, $p < 0.001$). The model performs well by looking into the distance distribution of the trips from the ground truth and the model output. The overall similarity results are consistent with the results of ODMs where the model performs the best in São Paulo followed by the Netherlands and Sweden. In all three regions, the model applied to the calibration dataset approximates the ground truth data slightly better than the one applied to the validation

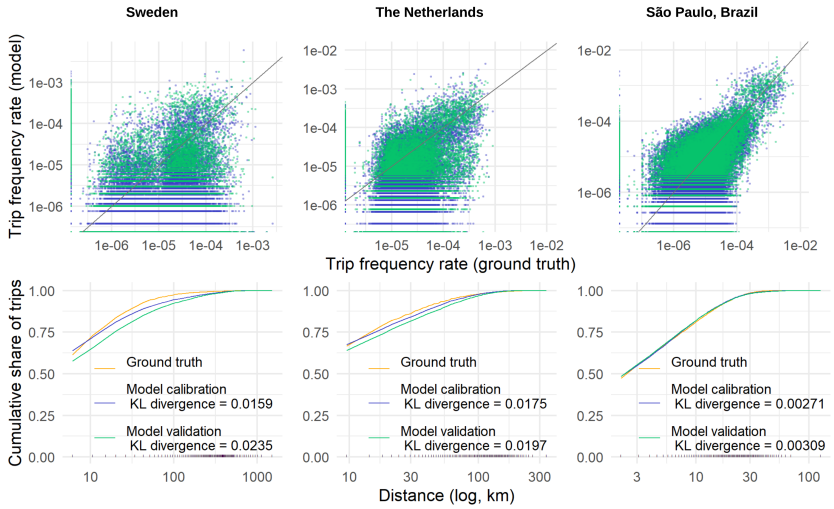


Figure 4.11: Top row: comparison of trip frequency rate for all origin-destination pairs between the ground truth data and model output where a data point represents a cell in the ODM. The blue dots are from the calibration results and the green dots are from the validation results. Bottom row: comparison of trip distance distribution between the ground truth data and model output. Source: Adapted from Figure 6 and 7 in Paper III.

dataset.

The model for São Paulo performs better than the ones for Sweden and the Netherlands. This discrepancy could be caused by how the selected Twitter users' home locations are distributed across the study area. Previous studies have suggested that most active Twitter users live in urban areas [10, 136] and that using sparse geolocations of Twitter data for simulating travel demand is more suitable for urban residents than for the population as a whole.

Characterising trip distance distributions

The trip distance distributions from the model-synthesised trips using sparse geolocations from 22 regions largely follow lognormal distributions and they reflect reasonable characteristics of regional heterogeneity.

After the experiment based on the ground truth data, we identify the optimal parameters for the model for Sweden, the Netherlands, and São Paulo, Brazil, respectively. We take the average values of the parameters from these three models and apply them to the 22 global regions for which we have geolocated Twitter data to create synthesised domestic trip distance distributions. We create the domestic trips generated in 140 simulation days. In order to test the capability of the proposed model, we characterise the

trip distance distribution of multiple regions with the synthesised domestic trips (black dots) shown in Figure 4.12). The lognormal model shown in blue curves approximates all regions well except for Manila, Philippines, which seems to follow the power law model with a heavy tail.

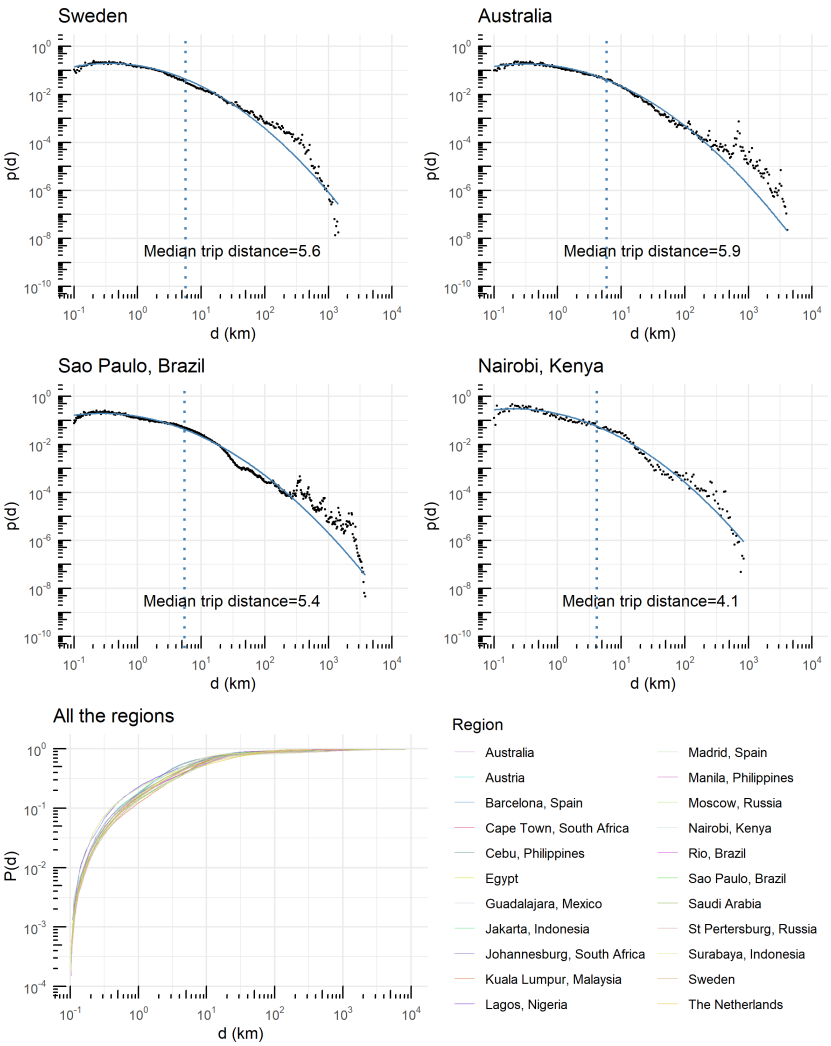


Figure 4.12: Selective distributions of synthesised domestic trips, $p(d)$. Black dots represent the probability density function. Blue curves stand for the lognormal/power-law curves fitted to the data. The last chart put all the regions together displaying their cumulative distribution function (CDF) of trip distance, $P(d)$. Source: Adapted from Figure 8 in Paper III.

We correlate the synthetic trip distance distributions with key regional characteristics. For instance, the domestic trips generated by city residents have more extreme values than those generated by the residents of the whole country. Not surprisingly, we find that the median trip distance correlates with the country area. We find that the higher the number of users, the longer the median trip distance. This suggests that the number of individuals in the model affects the distribution of synthesised trips. Without implying any causation, we recognise that these correlations should be examined more carefully with ground truth data in future studies.

Conclusions

In order to extend the use of the inexpensive and easy-to-access data such as CDRs, LBSNs, and social mdedia data, this study proposes an individual-based mobility model to fill the gaps in the sparse mobility traces for travel demand modelling. We validated our model and found good agreements on origin-destination matrices and trip distance distributions for Sweden, the Netherlands, and São Paulo, Brazil. The proposed model can be used to synthesise mobility at any geographical scale, and the results can later be applied to modelling travel demand. We further apply the model to characterise domestic trip distances for a mixture of cities and countries globally. The trip distance distributions from the model-synthesised trips using sparse geolocations from 22 regions largely follow lognormal distributions and they reflect reasonable characteristics of regional heterogeneity.

The proposed model for filling the gaps in sparse individual mobility traces has some limitations. The proposed model fills in the data gaps of individual mobility. However, due to the lack of matching individuals, our validation data represent the aggregated pictures of population flows and trip distances. More steps can be taken to address the inherent inconsistency between the proposed individual-based model and the calibration to the population data. One future direction is to test the performance of the proposed model using high-resolution GPS data: with a more complete set of mobility traces, we can simulate a variety of sparsity levels by downsampling the observed locations and evaluate the impact of sparsity on the model's performance. Further exploration is needed to understand the regional differences between the 22 cities and countries tested. Despite difficulties in obtaining good quality ground truth data, more validations would improve model credibility and usability. Last but not least, the temporal dimension can be added to the proposed model. The model can be extended in future studies by combining spatial and temporal dimensions so that the synthesised mobility data can be more useful in transport planning such as congestion management.

CHAPTER 5

Transport modal disparities between public transit and car

Using emerging data sources including Twitter data and many other GIS data, Paper IV-V reveal how public transit and car driving are different in terms spatiotemporal patterns of travel time and how the potential competition between ride-sourcing and public transit manifests. They attempt to answer the below question:

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

An overview of the research scope and the involved data sources are presented in Figure 5.1. Paper IV [13] and Paper V [137] combine multiple data sources to characterise transport modal disparities between car and PT, which further demonstrate the modelling potentials of the applied data sources.

Paper IV reveals the disparities in travel time between car and PT in four cities. A combination of multiple emerging data sources empowers a finer depiction of the spatiotemporal patterns than previous studies. The role of Twitter data is to provide the dynamics of travel demand. Therefore, Paper IV can be regarded as an application of Twitter data in real-world settings. Paper V explores the use of ride-sourcing trip data for understanding urban mobility, compared with PT as a potential alternative. Specifically, it looks into how the trip attributes and built environment are linked to the competition between the two modes and the implications for policymaking.

The following sections provide a summary of the appended papers on their motivations, research questions and methods, main findings, and conclusions.

5.1 Spatiotemporal patterns of travel time (Paper IV)

Disparities in travel times between car and transit: Spatiotemporal patterns in cities

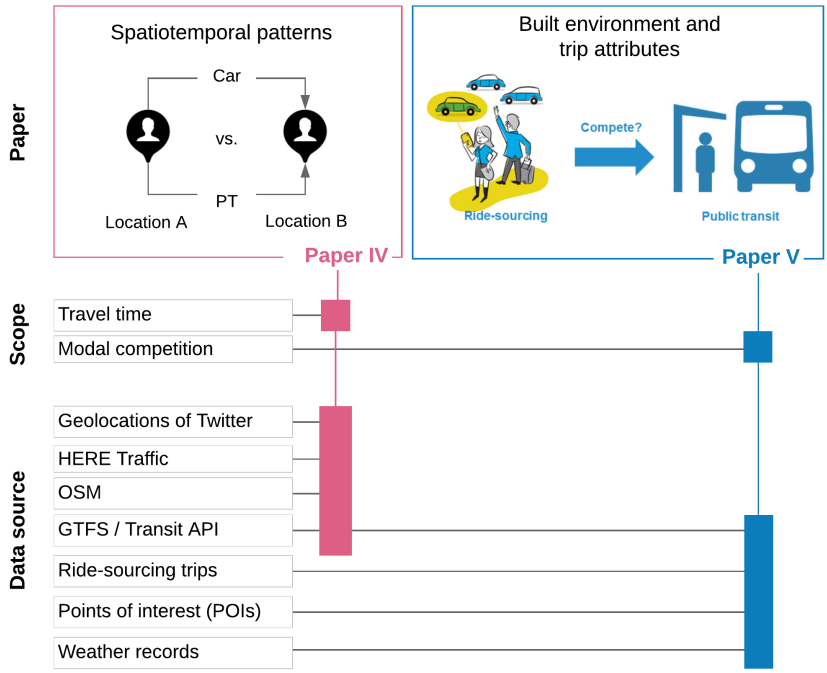


Figure 5.1: Overview of included studies: their scope and involved data sources.

Motivation

Many cities worldwide are pursuing policies to reduce car use and prioritise public transit (PT) as a means to tackle congestion, air pollution, and greenhouse gas emissions. The increase of PT ridership is constrained by many aspects, and among them travel time and the built environment are considered the most critical factors in the choice of travel mode.

The growing body of literature in understanding the spatiotemporal disparities in travel times for cars and PT [138, 139] starts using detailed spatial data and time-varying transport datasets, which provides opportunities for a more realistic assessment of modal disparity on travel time in this study. However, it remains to be explored how such disparity varies when considering the real travel demand. A full and realistic understanding of the disparities in travel times between these two modes could help identify opportunities of where and when public transit is competitive (time-wise) with automobiles and shed light on the relative transportation disadvantage of members of the community who must depend on public transit. Large-scale, representative dynamic travel demand data are critically needed for a more realistic assessment of this time disparity.

Research questions and method

Twitter data, specifically the density of geotagged tweets, reasonably capture an accurate representation of where and when people are engaging in various activities with high spatiotemporal resolution, therefore making it a good and low-cost source for obtaining dynamic travel demand in cities. This study leverages multiple large-scale data sources to capture, at a fine resolution, the spatiotemporal patterns of how car and PT travel times vary in four different cities: São Paulo, Brazil; Stockholm, Sweden; Sydney, Australia; and Amsterdam, the Netherlands.

Paper IV calculates the detailed spatiotemporal variations of travel times for an average weekday to improve the level of resolution at which we can understand the disparity in travel times between PT and car. We combine multiple data sources: HERE Traffic data over one year to derive empirical road speed, Twitter data accumulated from the past nine years, up-to-date GTFS transit data, and road networks from OpenStreetMap. Each city is divided into a hexagonal grid system, and travel times are estimated at different times of the day for any cell within the system (for more details, see Methods), calculating the door-to-door travel times by car and by PT to any highly visited cell (destination), identified as such based on geotagged tweet volumes. Within a selected time interval (e.g., 8:10 am to 8:25 am), the average travel time of a given origin cell is defined as the mean value of the travel times from that origin to multiple destinations whose volumes of geotagged tweets are used as weights. To quantify the modal disparity of travel time, we use the travel time ratio (R), defined as the travel time by PT divided by the travel time by car for a given origin-destination pair at a certain departure time. Finally, we visualise and analyse the results to demonstrate how car and PT travel times vary spatiotemporally across all the cities studied. Lastly, we present a systematic cross-regional comparison of the travel time disparity between car and PT in the four cities studied.

Main findings

Spatiotemporal patterns of travel times

Spatiotemporal patterns of modal disparities in travel times are shown in Figure 5.2. The travel time is the citywide average across departure locations, weighted by population density, of the average travel times from those locations. The shaded area indicates the range from the 25th to 75th percentile. Also shown is the percentage of grid cells accessible by PT by time of day. The inset figures are zoomed into the time period of from 05 hours to 23 hours to better show the variation of the travel time by PT. The value of the travel time ratio (R) for each cell as the origin is the average value based on the 5th to 95th percentile of travel times by PT and car in the time period between 05:00 and 23:00 weighted by the frequency of geotagged tweets in

the destination. The warmer the colour, the greater the advantage of car use over PT.

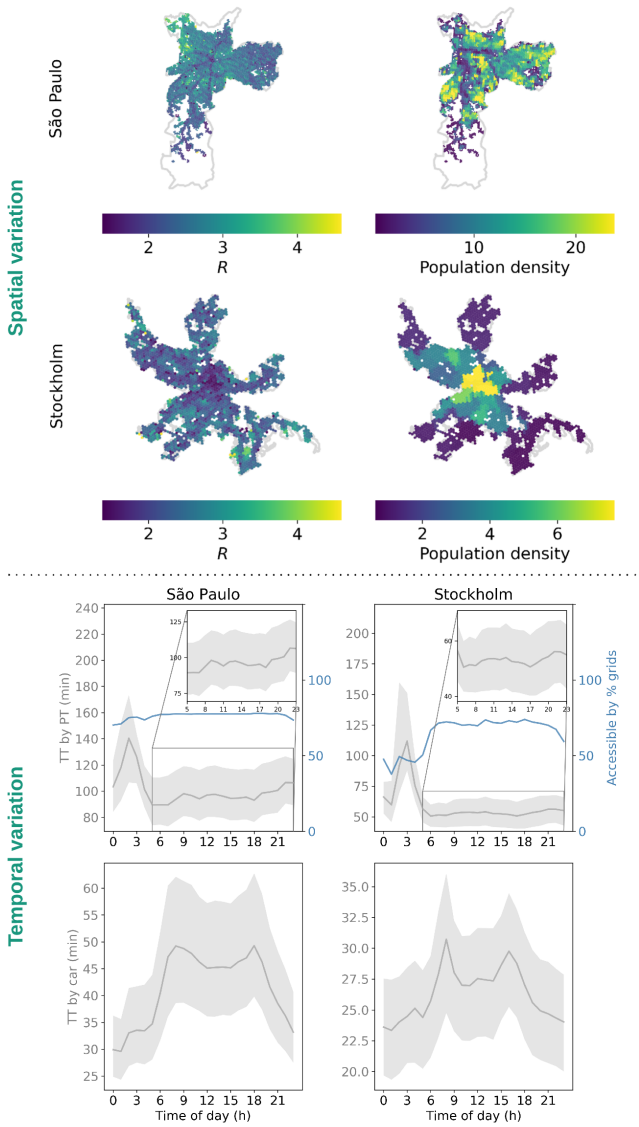


Figure 5.2: Spatiotemporal patterns of modal disparity in travel time for selective cities (adapted from Figure 1 and 3 in Paper IV). For Temporal variation, travel time by PT (upper row) and car (bottom row) are presented over the course of an average weekday. For Spatial variation, travel time ratio (R) to frequently visited locations (top row) and population density (bottom row) in 1000 persons per sq. km are presented.

The outcomes of the improved travel time calculations demonstrate the usefulness of applying large datasets in the framework developed in Paper IV. **It is shown how the travel time for each mode changes by time of day for an average weekday and how travel time varies spatially in different cities**. Future studies can zoom in and overlay infrastructure information to gain more detailed insights at the local level. This allows for urban planning policies to be better informed, especially in encouraging a mode shift from car to PT. While trips by PT take on average around twice as long as by car, this difference varies widely with location and time of day. **In general, the area in the studied cities where PT can outperform car use is very small, despite there also being substantial areas surrounding PT lines where the disparity of travel time by car and PT is smaller than in the rest of the city**.

Cross-regional insights into the modal disparity in travel time

The four cities have similarities and differences in terms of travel time ratio (R) as shown in Figure 5.3. **The average travel time ratio is around 2 throughout most of the day**, and the highest disparity between the two modes occurs between midnight and before dawn, when PT service is typically reduced or not running at all (Figure 5.3A). **The share of area that favours PT over car use is very small**: 0.62%, 0.44%, 1.10% and 1.16% (daily average) or 0.65%, 0.48%, 1.22% and 1.19% (during peak hours) for São Paulo, Sydney, Stockholm, and Amsterdam, respectively. In Figure 5.3(B), R **can be less than 1 (PT faster than car use) for distances < 3 km, but PT quickly loses the advantage as distances increase. Except for Stockholm, the cities show similar patterns when travel distances continue to increase: The disparity between PT and car travel time continues to increase until it reaches a maximum value at around 15 km, and then it starts to drop**. In addition, as shown in Figure 5.3(C), population density and R are also correlated: **The greater the population density, the lesser the disparity between PT and car travel times**.

Paper IV further summarises the city level performance of PT and car use in terms of the travel time ratio and the aggregate travel speed (Table 5.1). **At the city level (with grid cells weighted by population density), the lowest travel time ratio is observed in São Paulo, followed by Amsterdam, Sydney, and Stockholm in ascending order**. PT services in São Paulo and Amsterdam are more closely matched with where people live versus the PT services in Stockholm and Sydney, which are focused more on spatial coverage. For PT, the differences of (population weighted) speed are small across the cities. For São Paulo, the low driving speed suggests heavy traffic congestion, explaining why the disparity in time between PT and car is smallest there.

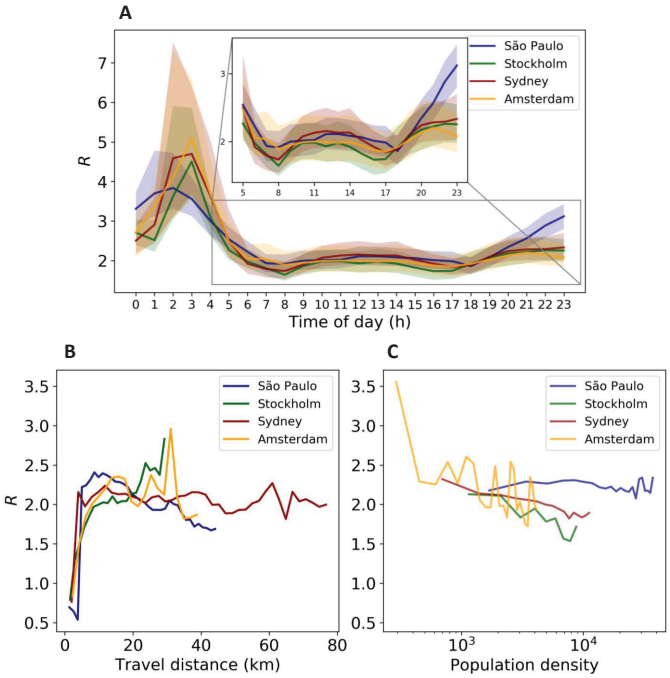


Figure 5.3: Travel time ratio across four cities (adapted from Figure 4-5 in Paper IV). (**A**) Temporal variation of citywide average travel time ratio (R). The shaded area indicates the two mid quartiles. The insert zooms in on the period from 05 hours to 23 hours, to better show the temporal variation of R. (**B**) Travel time ratio (R) as a function of travel distance. (**C**) Travel time ratio (R) as a function of population density. The unit of population density is 1 person per sq.km.

Table 5.1: Travel time ratio at the city level. [a–b] Average value weighted by population density in each grid cell. The travel time ratio at the city level is calculated based on the average value across all grid cells at all times of the day weighted by the frequency of geotagged tweets of the destinations.

City	R	R_{pop} [a]	Speed (km/h)		Speed$_{pop}$ [b] (km/h)	
			Car	PT	Car	PT
São Paulo	2.2	1.4	19.4	9.2	19.9	14.3
Stockholm	2.0	2.6	25.7	12.9	37.6	14.9
Sydney	2.2	2.3	33.8	16.6	30.9	13.9
Amsterdam	2.2	2.1	31.5	15.0	27.6	13.7

Conclusions

One significant contribution of Paper IV is the data fusion framework including real-time traffic data, transit data, and travel demand estimated using

Twitter data to compare the travel time by car and PT in four cities (São Paulo, Brazil; Stockholm, Sweden; Sydney, Australia; and Amsterdam, the Netherlands). The framework demonstrates its usefulness by revealing the travel time disparity between public transport and cars at a high spatial and temporal granularity enabling detailed and local-level explorations.

Moreover, Paper IV demonstrates that using PT takes on average 1.4-2.6 times longer than driving a car. The share of area that favours PT over car use is very small: 0.62% (0.65%), 0.44% (0.48%), 1.10% (1.22%) and 1.16% (1.19%) for the daily average (and during peak hours) for São Paulo, Sydney, Stockholm, and Amsterdam, respectively. The travel time disparity, as quantified by the travel time ratio R (PT travel time divided by the car travel time), varies widely during an average weekday, by location and time of day: there is less disparity near city centres, around PT lines, and during congestion hours. But R becomes extremely large ($R > 5$) at night when few transit services are available. A systematic comparison between these two modes shows that the average travel time disparity is surprisingly similar across cities: $R < 1$ for travel distances less than 3 km, then increases rapidly but quickly stabilises at around 2.

This study contributes to providing a more realistic performance evaluation that helps future studies further explore what city characteristics as well as urban and transport policies contribute to make public transport more attractive, and to create a more sustainable future for cities.

5.2 Modal competition (Paper V)

Ride-sourcing compared to its public-transit alternative using big trip data

Motivation

A core transportation strategy to mitigate negative environmental impacts is shared mobility, which refers to the services and resources involved in using a motor vehicle, bicycle, or other low-speed transportation mode that is shared among users, either concurrently or one after another [24, 140]. Public transit (PT) / mass transit and ride-sourcing (here, the latter refers to on-demand mobility services via smartphone apps to connect drivers with passengers) are both included in shared mobility. As a mode of shared mobility in cities, ride-sourcing services become increasingly popular; one of the key questions remains unanswered: Does ride-sourcing complement, or compete with, PT?

Though it is important to understand the interplay between ride-sourcing and PT, the potential for ride-sourcing services to replace PT trips has largely been overlooked [30, 86, 141]. The relationship between ride-sourcing and PT remains elusive, especially in developing countries where data are often lacking. Meanwhile, a rapidly growing body of literature uses advanced

techniques in machine learning and network science with big trip data to model urban structure. However, they remain under-exploited on revealing the impacts of trip attributes and the built environment on the relationship between ride-sourcing and PT.

Research questions and method

Paper V explores the conflict relationship (i.e., competition) between ride-sourcing and PT through the lens of big data analysis. For the characterisation of trip attributes and built environment, we incorporate functional urban regions identified using POIs with clustering analysis, transit access, and the community structure of ride-sourcing demand. We use a glass-box model that predicts whether a ride-sourcing trip directly competes with its alternative PT, providing intelligible outputs that can easily be visualised. The main factors include travel time for ride-sourcing and PT, weather condition, functional regions, transit access, and demand-based communities of pick-up and drop-off zones. Specifically, three nested questions are explored, as shown below:

- Does ride-sourcing compete with public transit?
- What trip attributes and built environment are linked to the competition?
- What are the implications for policymaking?

In order to answer the above research questions, Paper V develops a data-fusion framework, as shown in Figure 5.4. In preprocessing the original dataset, we first filter out abnormal request records and enrich each record with the travel information for its PT alternative assuming the same departure time, origin, and destination as well as with the weather information for the departure time. Moreover, we detect the community structure of the ride-sourcing origin-destination matrix created by connecting all the pick-up and drop-off zones. 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 the trend for the competition between them.

The original dataset consists of a series of records with the origins and destinations of ride-sourcing trips but without any informative 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 and quantify the transit access density in the zones. The descriptive statistics based on the processed trip attributes and the built environment of the ride-sourcing trips used for modelling are summarised in Table 5.2.

In the model construction, we first label the processed ride-sourcing trips as transit-competing or non-transit-competing based on the time of day

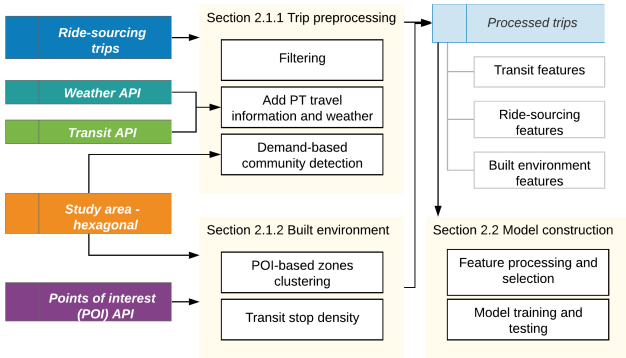


Figure 5.4: Methodological framework of Paper V. The arrows mark the flow of data. Source: Figure 1 in Paper V.

and the walking distance to and from transit stations: If a ride-sourcing trip, when instead served by PT, were to have had an access/egress walking distance of less than 800 m (each), and depart between 6 am–11 pm, it is called a transit-competing trip ($y = 1$), otherwise it is non-transit-competing ($y = 0$). Next, we process the features to eliminate multicollinearity and select the qualified features. Finally, we construct an enhanced GAM (as introduced in Section 3.1.2) to characterise the two categories in terms of the trip attributes and the built environment. Whether a ride-sourcing trip is predicted as transit-competing is dependent on the summation of the scores of the smooth function of features/feature interactions. Therefore, a single feature or feature interaction scoring above 0 increases the chance of the sample being transit-competing ($y = 1$), i.e., increase the tendency to be transit-competing.

This study explores the conflict relationship (i.e., competition) between ride-sourcing and PT through the lens of big data analysis. The contributions pertain to methodological and empirical aspects. Methodologically, we apply a data fusion framework without involving empirical PT trip records. Applying a glass-box model on the enriched ride-sourcing trip data provides a good overview of not only the main factors affecting the relationship between ride-sourcing and PT, but also the interactions between those factors; the latter is lacking in the literature. From the perspective of gaining new knowledge, data from developing countries are generally under-exploited to discuss the relationship between ride-sourcing and PT. The obtained insights of this study are useful to guide the local transport planning and they also contribute to an improved big picture of how global cities are experiencing ride-sourcing.

Table 5.2: Descriptive statistics of the ride-sourcing trips used for modelling. Mean for continuous variables and % for categorical variables. Source: Table 2 in Paper V.

Variable	Levels	Mean (SD) or %
TT by ride-sourcing, min	-	22.30 (12.81)
TT ratio excl. access/egress walking	-	1.81 (0.85)
Transit-stop density (pick-up zone), 1/km^2	-	12.38 (12.43)
Transit-stop density (drop-off zone), 1/km^2	-	12.66 (12.74)
Weather	Clear	0.9
	Clouds	42.2
	Haze	4.8
	Fog	1.4
	Mist	41.8
	Rain	8.9
Demand community (pick-up zone)	North	30.2
	South-West	38.7
	South-East	31.2
Demand community (drop-off zone)	North	30.8
	South-West	39.2
	South-East	30.0
# of boardings	1	57.7
	2	35.6
	3	6.0
	4	0.6
	5	0.04
	6	0.003
Functional cluster (pick-up zone)	Centre	55.2
	Centre-business	18.7
	Transition	11.8
	Residential-business	2.9
	Outer-residential	4.5
	Business-residential	4.0
	Rural	3.0
Functional cluster (drop-off zone)	Centre	54.9
	Centre-business	18.0
	Transition	12.1
	Residential-business	2.7
	Outer-residential	4.4
	Business-residential	4.3
	Rural	3.6

Main findings

The by-definition competition

Of the 4.27 million ride-sourcing trips shown in Figure 5.5, 48.2% compete with PT according to the definition above. Despite the binary simplification of the relationship, this number suggests that a considerable share of ride-sourcing trips can potentially be done by taking PT. Both categories of ride-sourcing trips have drop-off hot spots at the international airport and the railway station where we find that the willingness to take PT for long-distance trips is less affected by transit access. In other words, despite good PT access to the airport and railway station for those transit-competing trips, some travellers may be unwilling to take PT when, for instance, carrying luggage. One explanation for this is the consideration of vehicle comfort [142]; with a long journey by air or rail ahead, the passengers value the access trip more

than usual.

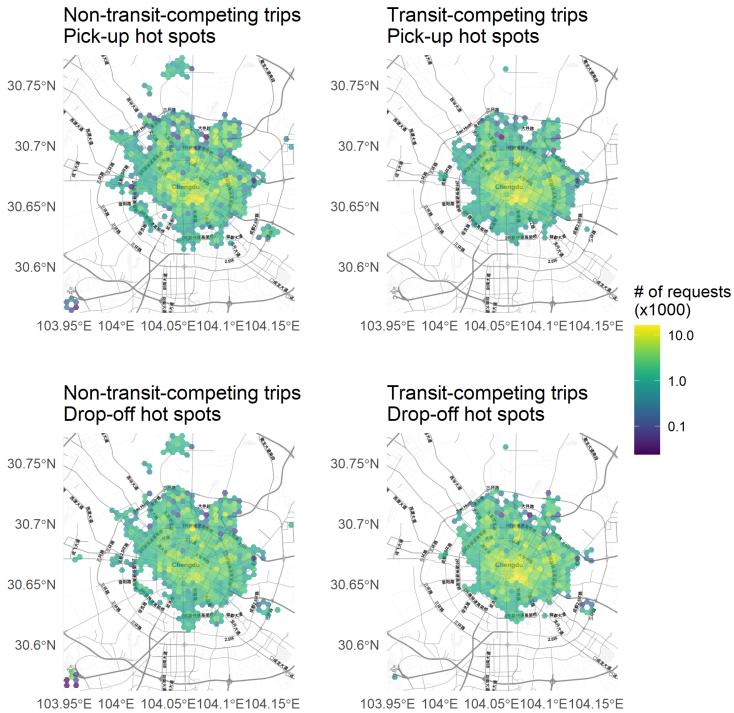


Figure 5.5: Hot spots of ride-sourcing trips. All the cells shown are statistically significant detected by Getis-Ord Gi* [135] with Z-score ≥ 1.96 and $p < 0.05$. They are hot spot cells at the 95% confidence level. The group of cells on the southwestern side are Chengdu Shuangliu International Airport. The group of cells on the southeastern side are a railway station. Source: Figure 5 in Paper V.

Trip attributes

The explored trip attributes interactively impose a significant impact on whether a ride-sourcing trip is transit-competing.

- **Travel time**

Short trips (< 15 min by ride-sourcing) tend to be transit-competing (Figure 5.6A). This can be explained by walking time being perceived negatively especially for short journeys [143]. For these transit-competing short trips, walking would take up a big share of total travel time were the trip done by PT. If one wants to ease the competition between ride-sourcing and PT for a

better mix of modes, this observation suggests decreasing the travel time by PT, especially for those ride-sourcing trips for which TT by ride-sourcing is less than 15 min.

Taking out the factor of walking, the TT ratio is an indicator used to reflect the in-vehicle time disparity between PT and ride-sourcing (Figure 5.6B). If the TT ratio is above 1.5, a ride-sourcing trip is more likely to be transit-competing. Given people's preferences for travelling faster where the PT alternative was on the slow side [142], the ride-sourcing service in the study area to some degree fills the PT demand gap where PT takes too long relative to ride-sourcing.

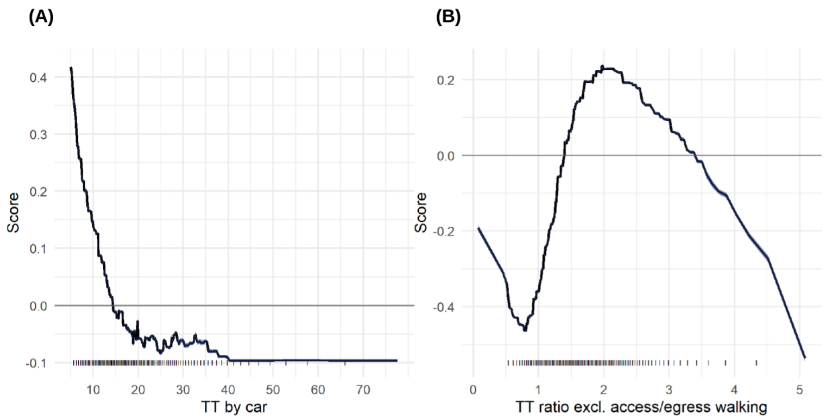


Figure 5.6: Impact of travel time. (**A**) Travel time by ride-sourcing. (**B**) Travel time ratio excluding access/egress walking. A score above zero means a tendency to be transit-competing ($y = 1$), while $y = 0$ for the score below zero. Black vertical lines indicate the value of ith percentile ($i = 1, 2, ..., 99$). Error bars show the 95-percentile confidence level of the score curve. Same below. Source: Adapted from Figure 8 in Paper V.

- **Transfer**

A study indicates that a transfer can be equivalent to 5 – 20 in-vehicle minutes [143]. The more transfers are needed, the more likely a trip is transit-competing (Figure 5.7A). This suggests that ride-sourcing covers the travel demand where the transfers are too many despite short access and egress walking distances. The in-vehicle time disparity between ride-sourcing and PT interacts with the number of transfers (Figure 5.7B): when at least one transfer is needed, the competition tends to happen even for those trips of little disparity. This highlights the penalty of transferring which makes PT less competitive than ride-sourcing. Therefore, the strategy to ease the competition can be decreasing the number of transfers.

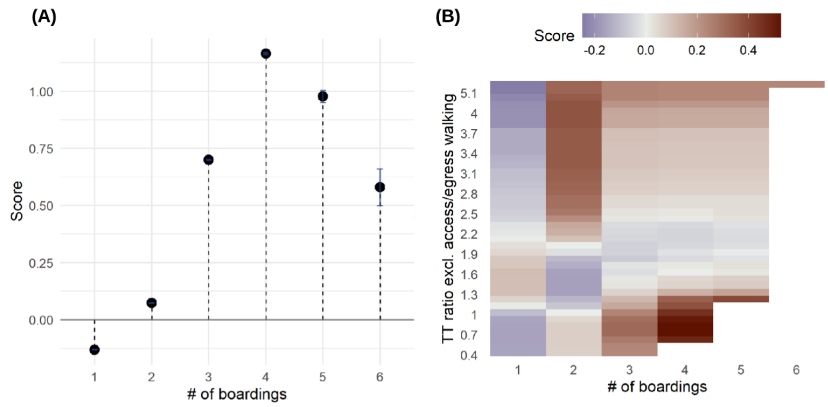


Figure 5.7: Impact of transfer. (**A**) Number of boardings (being 1 means no transfer). (**B**) Heat map of the score for the pairwise interaction component: # of boardings × TT ratio excl. access/egress walking. The blank areas have fewer than five ride-sourcing trip records, so the probability score is assumed to be unreliable. The areas coloured red and blue increase and decrease the probability of being transit-competing, respectively. Same below. Source: Adapted from Figures 8 and 9 in Paper V.

- **Weather**

Poor weather conditions such as fog, mist, and rain tend to ease the competition between ride-sourcing and PT (Figure 5.8A). This is consistent with previous findings that rainfall would typically increase the use of public transit as walking or driving might be quite difficult under such conditions [144]. It means that the willingness of taking PT increases under these weather conditions resulting in less transit-competing ride-sourcing trips. On the other hand, if the weather is clear, the competition tends to happen if the TT ratio is greater than 2.1 (Figure 5.8B). For the poor weather conditions, sensitivity to travel time disparity is low.

Built environment

The identified functional clusters and transit-stop density are used to characterise the built environment of the study area, which affects whether a ride-sourcing trip is transit-competing.

- **Functional clusters**

Seven functional clusters are created, see Figure 5.9. Cluster Centre is the first tier with the highest land-use intensity and diversity, which generates and attracts the most ride-sourcing trips. As the second tier, Clusters Centre-business, Transition, and Residential-business have a moderate level of land

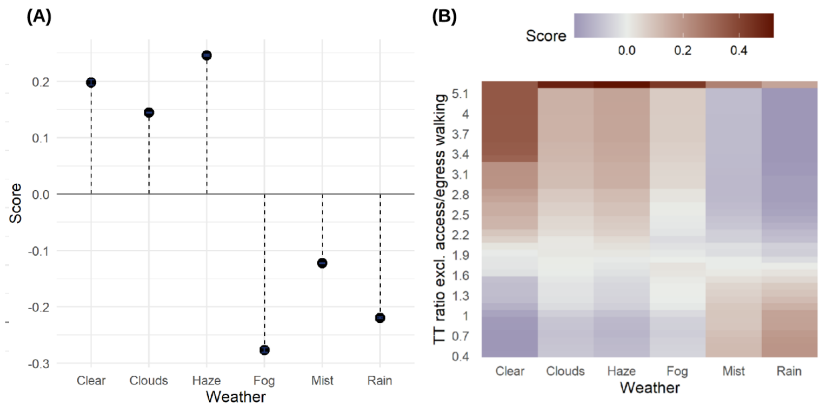


Figure 5.8: Impact of weather. (**A**) Weather. (**B**) Heat map of the score for the pairwise interaction component: weather × TT ratio excl. access/egress walking. Source: Adapted from Figures 8 and 9 in Paper V.

use, followed by Clusters Outer-residential, Business-residential, and Rural, as the third tier, in the transition area between the main city and the surrounding area. To take a closer look at these clusters, we define the share of commercial POIs per zone as the share of POIs of finance, beauty, life, shopping, hotel, and food, given these are POIs for the provision of goods or services. The share of commercial POIs of the clusters are Residential-business (52%), Centre-business (47%), Centre (44%), Transition (41%), Business-residential (31%), Outer-residential (29%), and Rural (21%) in descending order.

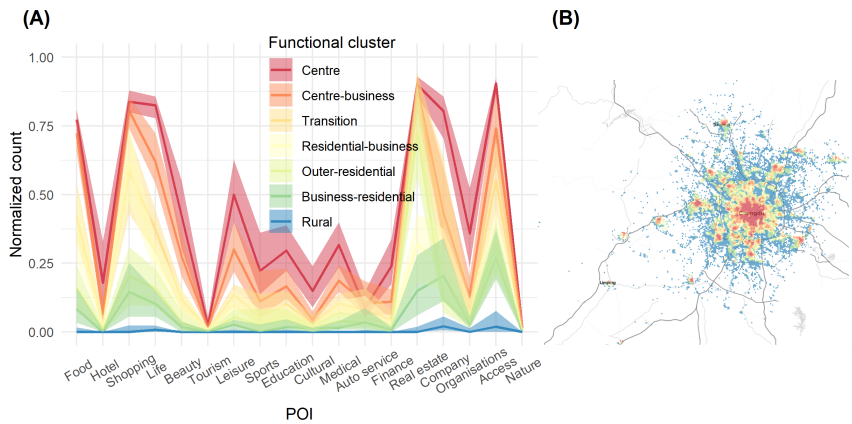


Figure 5.9: Functional clusters of zones. (**A**) The normalised number of POIs (of 18) in the functional clusters of zones. The shaded area indicates the range from 25th percentile to 75th percentile. (**B**) Spatial distribution. Source: Adapted from Figure 6 in Paper V.

Higher diversity and density of land use encourage the choice of non-driving modes [145]. However, despite the high diversity and density of land-use patterns in Cluster Centre (better access to PT as well), the ride-sourcing trips there have a slightly higher tendency to compete with PT (Figures 5.10A-B). On the flip side, the ride-sourcing trips in the other areas accounting for around 50%, such as Outer-residential and Rural where the land-use density/diversity is not as great as the central city, are less transit-competing. This implies that the role of ride-sourcing in Chengdu is leaning towards the complementary side to the PT system.

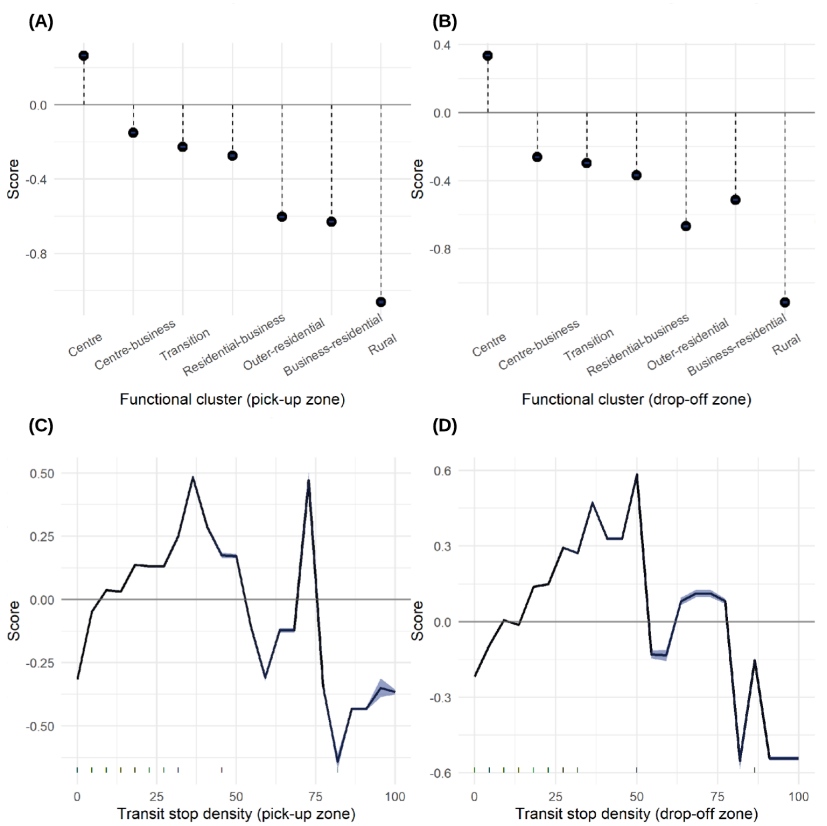


Figure 5.10: Impact of the land-use pattern and transit-stop density: single-feature effect. (**A**) Pick-up zone of the land-use pattern. (**B**) Drop-off zone of the land-use pattern. (**C**) Pick-up zone of transit-stop density. (**D**) Drop-off zone of transit-stop density. Source: Adapted from Figures 8 and 9 in Paper V.

The functional cluster also interacts with the other trip attributes such as with # of boardings (Figure 5.11A). Clusters Outer-residential and Transition are in the middle area between Chengdu city centre and the surround-

ing cities' centres, and they have a higher probability of generating transit-competing trips only if the PT alternative requires multiple transfers. These zones have a low density of economic activity according to their lower number of various POIs compared with the rest of the study area (Figure 5.9A). This suggests that there is room for improvement of PT in Clusters Outer-residential and Transition by reducing the transfer inconvenience by increasing connectivity between the central city and these areas.

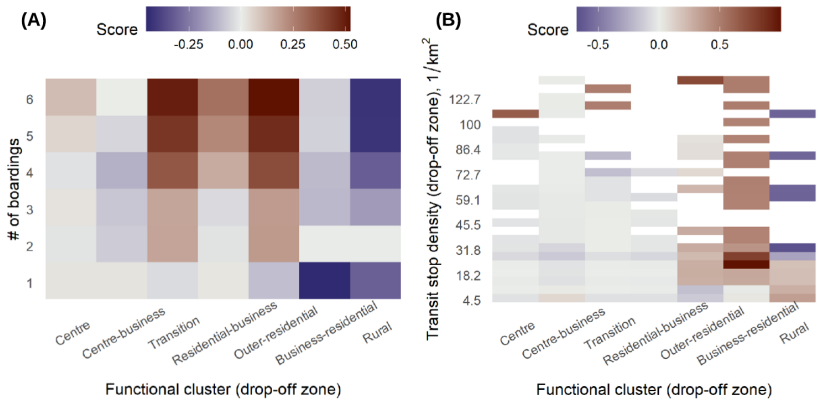


Figure 5.11: Impact of the land-use pattern and transit-stop density: heat map of the score for the pairwise interaction component. (**A**) Functional cluster (drop-off zone) × # of boardings. (**B**) Functional cluster (drop-off zone) × transit-stop density. Source: Adapted from Figures 8 and 9 in Paper V.

- **Transit-stop density**

We observe a negative impact of transit-stop density on the competition (Figures 5.10C-D); the better the transit access, the more likely a ride-sourcing trip is transit-competing. A similar relationship has also been found by [146], who find ride-sourcing services are not filling the demand gap in the areas of low transit-stop density. Moreover, the trips attracted to the zones of moderate to high transit-stop density in Cluster Business-residential are more likely to be transit-competing (Figure 5.11B). One explanation could be that given that Cluster Business-residential features a large number of business and much real estate, this tendency is due to a higher probability of business trips instead of private ones. As suggested by a previous survey study [147], the respondents who report higher numbers of long-distance business trips are also more likely to have used ride-sourcing services.

Conclusions

Spatio-temporally, the travel demand for transit-competing trips largely overlaps with that for non-transit-competing trips. The transit-competing trips account for 48.2% of the total trip records studied. Competition is more likely to happen when the travel time by ride-sourcing < 15 min or the travel time by PT is disproportionately longer than ride-sourcing (in-vehicle travel time ratio > 1.8). Requiring multiple transfers is also associated with the competition between ride-sourcing and PT, especially for the trips within the transition area between the central city and the outskirts. Poor weather conditions, such as rain, tend to ease the competition between ride-sourcing and PT, where the ride-sourcing users seem to be less sensitive to the travel time disparity between the two modes. Functional cluster of urban regions is the most important factor in determining the relationship between the two modes. Both low density and low diversity of land use are associated with a lower probability of generating transit-competing trips. The better the transit access, the more likely a ride-sourcing trip is transit-competing, especially for the areas featuring a large number of companies and real estate.

Some recommendations for transport planning based on the main findings are to: (1) Improve PT services that provide access to the international airport; (2) Expand PT networks guided by the transit-competing ride-sourcing trips featuring short travel time but a big gap between the two modes; (3) Increase the connectivity between the functional urban regions, Outer-residential and Transition, and the rest of the study area; (4) Incentivise the ride-sourcing trips that fill the gaps in the PT services that take a long time or require lengthy walking and transfers connecting to suburban areas; (5) Better combine the travel information of ride-sourcing and PT for travellers for the first- and last-mile issues.

CHAPTER 6

Discussion, outlook, and closing words

This chapter first gives an overview of the interconnections between the appended papers against the background of big data era. Next, Section 6.1 synthesises the findings in Paper I-III on the potentials and limitations of using emerging data sources, particularly geolocations of Twitter data for modelling mobility. And Section 6.2 discusses the findings in Paper IV-V on the transport modal disparities between car and public transit. Finally, this chapter reviews the limitations of the appended studies and points out the future directions.

Interconnections between the appended papers

The last decade witnessed a rapidly growing body of literature using new data sources with data fusion approaches in mobility and transport studies. The main drivers include, but not limited to:

- The ever-increasing availability of these emerging data sources and the ease of access to them.
- The increased cost of collecting traditional travel survey data together with the decreased response rate.
- The increased requirement of spatiotemporal resolution to enable better-informed policymaking and transport planning.

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 fundamental aspects of mobility (Paper I), a more systematic exploration of the data feasibility for travel demand estimation (Paper II), towards a more practical direction — addressing the identified issue of data sparsity with a new model for synthetic travel demand (Paper III). Involving more diverse data sources beyond geolocations of people's movements, the research continues with putting the movements of people into its context, transport systems; Paper IV-V provide the insights into the disparities between car and PT with

a high spatiotemporal resolution, which are useful for guiding the real-world practices, such as transport planning for PT and encouraging a modal shift from car to PT.

When using social media data for modelling mobility, the research started with the descriptive analysis using mobility metrics and models to reproduce the observed patterns in the previous studies in physics and transportation. After those efforts that generally found good agreement with the known universal patterns, the research gap has been narrowed down to a more practical direction: how to use social media data, e.g., to guide transport planning, and what improvements we need to make to the existing methods and data itself. With a deepened understanding of the pros and cons of social media data, the research continued with formulating the research questions that can be answered by using social media data.

Using geotagged Twitter data, mobility is firstly described by abstract metrics and physical models to reveal the population heterogeneity of mobility patterns (Paper I) and to estimate their aggregate travel demand (Paper II). We examine particularly the impacts of data sparsity, spatial scale, sampling methods, and sample size on the feasibility of using such a data source (Paper II). Aimed at addressing the identified sparsity issue in Paper II, we propose a mobility model that fills the gaps in the individual sparse traces to create synthetic travel demand characterising flows of people and their trip distance distribution (Paper III). These efforts improve understanding of modelling potentials and limitations of geolocations of Twitter data.

Human mobility in space and time (e.g., being observed from Twitter) is strongly influenced by ambient transport systems as well as travellers' housing conditions, socio-demographics, and travel preferences, etc. Besides the abstract representation of mobility using emerging data sources, a natural step forward is to put it into its context, transport systems. The other focus of this thesis is to improve the understanding of transport modal disparities between car and PT, in order to better inform the policy attempts to encourage a modal shift from car to the other low-carbon modes like PT.

Increasingly sensed GIS data in urban systems provide observations of transport systems at an unprecedentedly high spatiotemporal resolution. In this thesis, these data sources are combined together with Twitter data to understand the two aspects of modal disparities between car and public transit: spatiotemporal patterns of travel time (Paper IV) and modal competition (Paper V). We combine the travel demand as revealed by Twitter data, transportation networks, and historical road speed records, at a high spatial and temporal granularity. Such a data fusion framework presents a more realistic picture of the modal disparity in travel time between car and public transit in four cities in different countries. As travel time is only one aspect of the modal disparities, we further take built environment and other trip attributes into consideration to explore the modal competition between an emerging car mode, ride-sourcing, and PT. We develop a reproducible

framework of using open sources to enrich an incomplete but big-volume trip dataset and apply a glass-box machine learning model for intelligible outcomes. This framework reveals the spatially explicit unmet PT demand.

6.1 Potentials and limitations for modelling mobility

Trade-offs between data availability and biases

One key strength of social media data is the low cost and easy access when compared to the traditional data sources, which make it widely available. However, this low cost comes with a price: a biased population representation and low and irregular sampling of the actual mobility.

A **biased population representation** is confirmed by Papers I and II: top geotag Twitter users have clear signs of overly representing big-city residents. We apply a simple weighting method to adjust for this population bias; the ratio of Twitter users to the true population in the municipality of Sweden is given to each user's records when aggregating the results [4]. However, the method needs further improvement to select the appropriate spatial units for de-biasing such a data source.

Geolocations of Twitter data are from an **irregular sampling** of the actual mobility trajectories. We identify a clear pattern of leisure activities from the temporal distribution of the geolocations of Twitter data (Paper I). This suggests that the geotag users selectively report their locations. Tasse et al. [65] find that the geotag users tend to report uncommon places outside their daily routines. This creates an irregular distortion of the actual mobility where the most visited locations (e.g., home and workplace) are under-represented in Twitter data.

However, we can indirectly infer the rough location of a given user's home because people are habitual animals with a high regularity of daily movements, which is bounded by a limited time budget for travelling [148]. Therefore, a widely used approach is to assume that the most visited location during weekends and 7 pm - 8 are on weekdays is the home location and the second most visited location during 8 am - 8 pm on weekdays is identified as one's workplace [149]. We found that the locations with these two distinct temporal signatures widely exist in the geolocations of Twitter data (Paper I). Despite home and workplace being observed, they are under-represented in data. We conclude that this under-representation of home and workplace reduces the feasibility of using this data source to represent **commuting travel demand** as compared with the travel survey (Paper II).

As for the **low sampling**, geolocations of Twitter data are the sparse observations of the actual mobility. Even for those top geotag Twitter users in 23 global regions who geotag their tweets more frequently, the number of

reported locations per active day per user varies between 1.4 and 3.2 [12]. This number is below the usual number of visited locations per day, e.g., 3.14 revealed by the Swedish national travel survey [12]. Moreover, these users tend to have a long duration of not reporting any locations between active days. Consequently, geotagged social media data only capture incomplete mobility trips because they do not record all the locations a user has visited.

Potentials at individual and population levels

Despite the clear signs of the above limitations, **mobility regularity**, **diffusive nature**, and **returning effect** are preserved in the geotagged tweets to some extent (Paper I). Moreover, the fundamental patterns of **population heterogeneity** on mobility are reflected as well. However, these agreements on the fundamental patterns of human mobility do not naturally guarantee that we can use geolocations of Twitter data for a more practical use, e.g., travel demand estimation as quantified by ODMs.

This thesis further concludes that geotagged tweets contribute to a reasonably good travel demand estimation with stability over time. However, the potential for estimating travel demand using geolocations of Twitter data is affected by many aspects.

The main obstacle of using Twitter data for travel demand estimation on a large **spatial scale** is **data sparsity**. Given the overly representation of city residents, Twitter data is more suitable for approximating urban travel demand than the one at the national level.

Regarding the **sampling method** of geotagged tweets, collecting more detailed, long-term individual data from user timelines for a small number of individuals (the User Timeline API) produces more accurate results than short-term data for a much larger population within a region (the Streaming API). Therefore, the User Timeline API is recommended as opposed to the Streaming API due to the following reasons: (1) More time-efficient collection of a large number of tweets; (2) Longer-term observations of the covered individuals that compensate for incomplete mobility traces; (3) Better performance in approximating travel demand.

As for the impact of **sample size**, the minimum number of geotagged tweets for a reasonable travel demand estimation is explored in this thesis. We consider both the similarity between Twitter-based estimation and the ground truth as well as the stability of such similarity. We find a magnitude of 1,000 geotagged tweets is sufficient for the city-level (Greater Gothenburg, Sweden) and the national level (Sweden) requires 10,000 tweets to reach a stable similarity. However, how this finding fits the other regions needs further examination.

Another strength of social media data is the dynamics it naturally contains about where and when people do various activities, i.e., the spatiotemporal patterns. The stream of Twitter data continuously depicts the individuals'

activities in space. These dynamics help to create a more vivid picture of mobility at both individual and population level. Tasse et al. [65] suggest that most geotag users geotag their tweets within an hour of arrival (if at all), thus geotagging may be a timely indicator of the start time of the activity. Therefore, the density of geotagged tweets naturally reflects the attractiveness of places in cities. This density map can be applied to represent the attractions of places when modelling travel demand (Paper II). And it can also be used as a proxy for destinations when evaluating the travel time by car and PT (Paper IV).

Extending the use by innovative approaches

One issue of geolocations of Twitter data, caused by low sampling, is to what extent the covered mobility traces is incomplete, i.e., **sparsity issue**. The more complete they are, the better picture of human mobility they provide. Given the sparsity of this data source, the feasibility of using them is limited. One can extend the use of such an inexpensive and easy-to-access data source by overcoming this issue. To do so, this thesis develops two innovative approaches.

A **density-based approach** is proposed to increase the amount of geotagged tweets in modelling travel demand (Paper II). A common practice of modelling ODMs using Twitter data is to create trips by adding a time threshold to the geotagged displacements, where only 20–35% of data are used. The newly proposed model uses geotagged tweets as attraction generators as opposed to the commonly adopted trip generators when estimating the population flows. This significantly increases usable data, resulting in a better representation of travel demand measured by both ODMs and trip distance distributions. The success of this approach can be ascribed to the key assumption: **the generated trips between zones by modelling are proportional to the population and the number of activities of which are geotagged**.

A mobility model is proposed to fill the gaps in sparse mobility traces, particularly traces from geolocations of social media data, from which one can later synthesise travel demand (Paper III). The proposed model applies the fundamental mechanisms of exploration and preferential return to synthesise mobility trips [1]. However, the details of these mechanisms are designed to accommodate the sparse individual traces of geolocated social media data. The proposed model can be used to synthesise mobility at any geographical scale. We find that the model-synthesised trips approximate the ground-truth data well in the selective regions. The trip distance distributions from the model-synthesised trips using sparse geolocations from 22 regions reflect reasonable characteristics of regional heterogeneity.

6.2 Characterising transport modal disparities

Data fusion approaches

A better understanding of transport modal disparities calls for innovative ways of utilising different data sources, especially increasing amount of incomplete but big datasets are made publicly available such as mode-specific trip data.

The data fusion used in Paper IV is a novel approach that allows us to combine both transport service demand and operations while getting more granular results. Especially with geotagging being a timely indicator of the start time of the activity, the method uses Twitter data as a proxy for time-varying travel demand provides. This is different from other approaches such as accessibility-based analysis that focuses on fixed points travel time or travel time to places of important functions (e.g. workplaces), or average demand without temporal resolutions such as an ODM output from static models. Due to the easy access of geotagged tweets globally, this application can be generalised to multiple regions.

The openly available mobility data can cover trips from ride-sourcing, ride-sharing, taxi, e-scooters, and shared bikes. 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 [e.g., 30, 137]. To gain insights from using these data, one needs a data fusion framework to enrich the original dataset. This thesis demonstrates an example of how single-mode trip data can be enriched to better understand the modal competition between multiple modes (Paper V). The study uses weather API, transit API, and POI API to retrieve more data based on the locations and departure time of the ride-sourcing trips. After enrichment, the ride-sourcing trip data have their corresponding transit travel information, built environment characterised by POIs, and weather condition. In this study, due to limited data availability, commercial APIs are used to get transit information and POIs. This is not as flexible as using GTFS data or land use data from open data portals. In the future, more efforts are needed to make these data publicly available, especially in developing countries.

Besides the data fusion framework, a glass-box model enhanced by machine learning techniques is constructed with the enriched ride-sourcing trips to discuss the relationship between ride-sourcing and its PT alternative. This model is particularly beneficial for high-dimensional analysis. As shown in Paper V, how the many factors of trip attributes and built environment affect the competition between ride-sourcing and PT are quantified additively together with the impact of the interactions between them. And these intelligible results are clearly visualised.

Travel time and modal competition

Travel time is one of the most important factors in mode choice. Its high spatiotemporal depiction shows that using PT takes on average 1.4 – 2.6 times longer than driving a car (Paper IV). The share of the area where travel time favours PT over car is surprisingly small at a magnitude of 1%. This means that with the same origin and destination, PT is virtually always slower than car.

When comparing car with PT, this thesis finds that PT outperforms cars on shorter trips below 18 min (Paper IV). However, when it is ride-sourcing (by car), people mostly take short ride-sourcing trips (Mean = 22.3 min, Paper V). And for those short trips, their PT alternative usually does not have much advantage in terms of travel time; TT ratio is 1.8 for in-vehicle travel times and roughly 2.2 for overall travel times. How PT outperforms car for short trips is one story, but in what cases people choose to travel by ride-sourcing instead of PT is another story. This contrast is thought-provoking. In other words, the ride-sourcing covers a significant share of the short trips where PT takes disproportionately longer travel time than car.

One review comment challenged the necessity of the work about the travel time disparity between car and public transit (Paper IV): "...we already know that transit is slower and in the core of cities it's faster due to congestion and in the nights when there is no service car is better. So why do we need a paper to compare this across cities and tell us what we know?" It raises a critical question, what does spatiotemporal analysis add when the aggregated results confirm common sense? The spatiotemporal details provide nuanced insights that are helpful to identify gaps and opportunities by policymakers and planners, i.e., it reveals when and where the gaps and opportunities manifest. The analysis of the fine resolution of space and time reveals (Paper IV), on the one hand, how the gap between the two modes varies widely across cities and on the other hand how the effects of travel distance and population density are incredibly consistent at the meta-level. In comparing ride-sourcing with its PT alternative, we present both aggregate statistics about the share of trips that can be potentially replaced by PT while more importantly, we show when and where these trips tend to happen and these help decide whether there is unmet PT demand for transport planners to ease the modal competition, especially for these short trips by ride-sourcing.

Making public transit competitive

Given the significant emission benefits of taking PT instead of ride-sourcing or car in general, making PT competitive becomes important. To this end, ride-sourcing trips tell us how PT could be improved to be more attractive to travellers. The answer to whether there is a competition between ride-sourcing and PT depends on the definition of such a competition. However, similar to the suggestions from previous studies, our study found that a large

share of ride-sourcing could have been done by PT with a reasonable walking distance.

In order to promote the use of PT, one could decrease the travel time by PT, especially for those ride-sourcing trips for which TT by ride-sourcing is less than 15 min. Besides relatively long travel time by PT, this thesis also looks into more diverse factors that could make PT more competitive compared with driving (Paper V). For instance, one could decrease the number of transfers by increasing the connectivity between frequently connected pairs of zones, especially for the trips connecting outer rings of the cities. This study also identifies which sub-regions in the study area (Chengdu, China) tend to generate transit-competing trips. Putting the typical transit-competing trips identified by the model on maps yields local insights for transport planning towards a more competitive PT mode in the study area.

In practice, there are many aspects to make PT competitive. High-quality PT services are perceived as reliable, fast, easy to access, comfortable, affordable, etc, and they are designed to serve the demand well. Qualitative research plays an important role in understanding these aspects of PT comparing with car driving [150]. Quantitative research driven by emerging sources such as smart card data provides direct insights into PT ridership [151] and demand patterns [152]. Recent studies reveal the importance of comparative analysis using data from different modes to urban and transportation research [153]. Because this allows us to understand the interplay between different transport modes, especially the unwanted competition between emerging modes like ride-sourcing and PT. This thesis shows how different data sources can be combined together to gain insights into such competition without directly using PT trip data. This perspective of analysis and its methodological framework contribute directly to transport planning to make PT competitive by filling the identified unmet demand.

Outlook

The use of emerging data sources in mobility and transport has gone through the exploratory stage, towards the application side from Paper I to V. My doctoral research started with the hammer of data science without much knowledge about mobility or the interconnections between disciplines of this field. After four-year research, compared with the breadth and the potentials for much more knowledge needed by this field, this doctoral research is just a start. Not constraining the research into specific data sources, I believe the future work should ask relevant questions with powerful and innovative tools. Here, I highlight four potential directions to pursue.

(1) Extending the use of social media data for mobility modelling.

This thesis focuses more on the sparsity issue of social media data, however, their inherent population and behaviour biases are not sufficiently explored and addressed. This thesis only demonstrates the first step of a simple

method for the population bias issue: we relate the socio-demographic information of the Twitter users to their identified home locations. We need to further examine the reliability of this method for a better population sample from Twitter users. Future research needs to continue the efforts of **de-biasing** the data source so that its use can be further extended. For instance, deep learning techniques point to a promising direction where we can do the joint classification of age, gender, and status of social media users followed by a post-stratification to create a more representative population sample [154].

To date, daily and short-distance trips have been extensively studied by transportation and geographic researchers using traditional household travel surveys. However, the one-day travel diary and the long-distance travel survey module are typically conducted separately in the household travel survey. There is a tendency to underestimate the long-distance travels of which the patterns and frequencies are often poorly characterised. Given that found geolocations of Twitter data overly represent leisure activities and the activities that happen outside the routine mobility range, future research can use these data for modelling **long-distance travels** including tourists' travel behaviours and international travels. This is a direction that my doctoral research did not have time to pursue.

Despite having shown the clear advantage of representing the dynamics of urban activities, this thesis evaluates geolocations of Twitter data mainly from a static perspective. One future direction is to test such a data source at different levels of temporal resolution. For instance, the **temporal dimension** can be added to the proposed model in Paper III so that the synthesised mobility data can be more useful in transport planning such as congestion management.

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. One could combine the geolocations with the tweet contents for inferring trip purposes. This work can contribute to a more robust and more informative estimation of mobility patterns and travel demand. Using text mining and cutting-edge techniques of natural language processing, a large body of literature uses tweet content for travel behaviour and attitude research. Due to limited time, this direction is beyond the scope of this thesis but worth exploring in the future.

(2) Generating global synthetic mobility data for improving travel demand projections.

Towards the end of this doctoral research, the studies have become more cross-disciplinary and practical. What we just started but haven't had any results to put in this thesis is the attempt to integrate the mobility studies using emerging data sources into the energy systems' modelling for the transport sector. This contributes to reducing the carbon emissions in the transport sector by providing a better estimation of travel demand at the

global level, and therefore, more reliable projections into the future scenarios.

Compared with empirical mobility data, synthetic data, e.g., those based on geolocations of Twitter data, has the advantage of bridging the gaps in the data and avoiding privacy issues in the availability of commercial-in-confidence data. This thesis gives an example of synthesising trips from sparse geolocations of Twitter data and how the proposed model can be used across many global regions. A continuous effort will work on a framework of creating synthetic data that combines multiple data sources, including conventional data sources such as travel surveys and population censuses, infrastructure and land use data, and transport service data such as ride-sharing, taxi, e-bikes, smart cards, e-scooters, and bike-sharing.

This **framework of creating synthetic mobility data** will provide data and insights into travel demand that are compatible with existing energy systems' modelling approaches. Along this direction, innovative data fusion methods are required to deal with missing information, e.g., transport modes, inconsistent data sources and data format. We also face the issue of missing data for certain regions, and to make it cover as many regions as possible, we need to solve the problem of how to extrapolate the missing regions' travel demand projection based on the best knowledge of the regions where the data are more available.

(3) Combining multi-modal trip data for reducing transport carbon emissions.

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

This thesis has demonstrated that HERE Traffic data, OSM, and GTFS data provide rich information for exploring the carbon emission reduction potentials in cities at a fine spatial granularity. With these data sources plus ride-sourcing trip data without the concurrent PT trip data, this thesis has asked the question in a virtual space: "What if these ride-sourcing trips were done by taking PT?" This "what-if" perspective lacks the capability of gaining insights into how different transport modes are adopted by travellers in reality. To better inform policymaking, PT big trip data need to be collected from other sources, for instance, smart cards, combined with the other emerging transport modes such as ride-sharing, e-scooters, and bike-sharing. They help to answer the questions, e.g., how are ride-sourcing trips and car trips distributed spatially? Are they different? How largely do they overlap?

Increasing number of studies focus on the potential of emerging transport modes, e.g., bike-sharing and e-scooters, for reducing carbon emissions. It is oftentimes naive to assume that a car trip below a certain distance can be replaced by bike-sharing or e-scooters while ignoring their use scenarios and consumer perceptions; though a simple assumption to evaluate such potential provides at least an upper limit of how much these emerging modes may reduce carbon emissions. On the one hand, combining the data of actual

car trips and trips that are done by emerging modes tells us how the reality deviates from its potential. And a big divergence suggests a big space for policy action. On the other hand, we need to combine user preferences and empirical trip data to reach a more realistic evaluation of different transport modes. The good news is that more and more cities have started their smart city initiatives to increase the data availability to the public. Empowered by these data sources, one of my future research directions will be studying **the occupancy, shareability, and electrification of new mobility services** provided by transportation network companies (TNC).

(4) Introducing the perspective of networks into urban mobility research.

Aggregating trips at individual or population level yields a network representing the interactions between places. This network at the population level is equivalent to ODM in this thesis. A series of metrics and models from network science has been used in this thesis to quantify the patterns of such spatial interactions between places generated by different groups of individual travellers (Paper I) or a population by certain transport mode (Paper V). However, the perspective of networks has only been implicit in this doctoral research, and the use of network science tools has been superficial and practical. Complex systems kick-started the theoretical education of my doctoral research, where network science was one of the modules. The early efforts of using the perspective of networks in mobility study were limited to descriptive analysis based on the networks of individual mobility trajectories and population flows. They have never been published probably due to unclear research questions.

Nevertheless, 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)** which tells us about social segregation and how such segregation and spatial interactions shape each other. This requires to relate one network to another, for which bipartite graph is a good tool. Moreover, the recent developments in deep learning for graphs, graph embeddings and graph convolutional networks (GCNs), allow us to directly learn and predict at the network level. These advanced techniques will contribute to urban mobility research together with the increasingly available data.

Closing words

At the end of this thesis, I'd like to first reflect on the data privacy issue which is not directly related to my research but an important topic for any researchers who use privacy-sensitive data in their work. And then I will present a general reflection on the future of the research field.

(1) Privacy and personal data protection

The issue about privacy and personal data protection concerning the use big data has gained increasing attention. The European Union has introduced a new privacy law, General Data Protection Regulation (GDPR), bringing new restrictions on how personal data is collected and handled. Sometimes personal information can be easily inferred due to extractable behaviour patterns that leak rich clues, especially they have easy access to high-quality, massive, and valuable personal data. For instance, we have shown that we can guess where a particular user lives from his/her geotagged tweets.

On the flip side, a trend in academia is to increase the publication reproducibility; it is becoming prevalent that researchers provide a repository of the data relevant to their publication. Without careful processing, the data could harm personal privacy. Although particular methods of anonymisation could be done to protect privacy, the publicly accessible data set could still be different from the one that researchers' publication is based on.

In the doctoral research, we apply public data from the Twitter platform. Every tweet collected has obtained the user's consent. However, the data can be potentially used to tie back to people of actual identity. As a researcher using such public data, one should not breach any personal data in the publications that can tie back to any users who contributed to the collected data. One should not put raw data into public without doing anonymisation. I believe researchers should also avoid privacy probing out of the researcher's own interest during the whole process of researching.

Specifically, in the storage of data, we guarantee a limited number of people can have access to it. As for the paper writing, we make sure that the results are presented at a certain aggregate level without showing individual physical coordinates or any traceable clues. Oftentimes we need to anonymise GPS trajectories properly before publishing or data sharing. There is a rapidly evolving research field about the techniques of anonymising for privacy protection as well as privacy-preserved models for mobility modelling.

We need a new framework to address ethical issues brought by using big data, without harming the process of increasing the reproducibility of doing science. We should actively interact with decision makers rather than being satisfied by following the bottom line of privacy protection in the current situation. Moreover, we need to know how to deal with big data, and at the same time, to have the sense of mission to protect personal privacy and to avoid unintended harm to society.

(2) Mobility data science

Mobility (human mobility) studies the large-scale movements of people in space and time. This thesis seats at mobility data science where data science techniques play a major role in the exploration of people's movements. This thesis has been mostly descriptive, i.e., taking care of "what" and "how" questions, due to the practice-driven perspective and the central role of data and its potentials and limitations. Therefore, most of the presented

work is deductive research. However, the scope of this field is far broader than what has been presented; it explores the equity issues in transportation, emerging mobility modes, how technologies affect mobility, why urban systems emerge in their certain structures, etc. Especially the recent COVID-19 pandemic has created a huge body of literature on human mobility and the pandemic spread/coping.

In the future, mobility data science will become more active because of 1) more available data of people's movements and transport systems, 2) mobility being the core manifestation of the interactions between people and their living environment, and 3) better-informed policymaking for coping with challenges of sustainable development.

It has been only recently possible to access people's movements at a large scale, while the inquiry of universal laws of mobility traces back to 1885 where human migration was represented with gravity models [155]. Nowadays, this idea is still being used widely in mobility studies. The mobility research driven by data should be rooted in theories to provide applicable knowledge for engineering. On the other hand, these emerging data should also contribute to developing better theories, when certain fundamental principles are found not to be universal after being more extensively tested by those data.

We have witnessed the trend that methods across different disciplines are being brought in for enhancing the understanding of human mobility, therefore, more cutting-edge methods from data science and complex systems will be applied to mobility data science in the future. For example, this thesis has demonstrated that the recent development in interpretable machine learning can be used to understand the modal competition between ride-sourcing and PT (Paper V). And the individual mobility model proves to be useful in generating mobility data with sparse input (Paper III). One of the future directions in Outlook also points to the deep learning techniques for graphs for studying mobility from the perspective of networks.

Most important, mobility data science will answer more research questions to solve societal and environmental challenges. For example, during COVID-19, researchers have been predicting pandemic spread and validating related policies guided by mobility data and studies. The community will continue with more efforts dealing with climate change challenges, natural hazards, social segregation issues, etc.

References

[1] C. Song, T. Koren, P. Wang and A.-L. Barabási (2010a). Modelling the scaling properties of human mobility. *Nature Physics* **6** (10), p. 818.

[2] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton and S. H. Strogatz (2010). Redrawing the map of great britain from a network of human interactions. *PloS one* **5** (12), e14248.

[3] V. Belik, T. Geisel and D. Brockmann (2011). Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X* **1** (1), p. 011001.

[4] Q. Wang, N. E. Phillips, M. L. Small and R. J. Sampson (2018). Urban mobility and neighborhood isolation in america's 50 largest cities. *Proceedings of the National Academy of Sciences* **115** (30), pp. 7735–7740.

[5] W. Huang, S. Xu, Y. Yan and A. Zipf (2019). An exploration of the interaction between urban human activities and daily traffic conditions: a case study of toronto, canada. *Cities* **84**, pp. 8–22.

[6] IPCC (2013). *Climate change 2013: the physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, p. 1535. ISBN: ISBN 978-1-107-66182-0. DOI: 10.1017/CBO9781107415324.

[7] Y. Yue, T. Lan, A. G. Yeh and Q.-Q. Li (2014). Zooming into individuals to understand the collective: a review of trajectory-based travel behaviour studies. *Travel Behaviour and Society* **1** (2), pp. 69–78.

[8] C. Song, Z. Qu, N. Blumm and A.-L. Barabási (2010b). Limits of predictability in human mobility. *Science* **327** (5968), pp. 1018–1021.

[9] T. H. Rashidi, A. Abbasi, M. Maghrebi, S. Hasan and T. S. Waller (2017). Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges. *Transportation Research Part C: Emerging Technologies* **75**, pp. 197–211.

[10] Y. Liao, S. Yeh and G. S. Jeuken (14th Nov. 2019). From individual to collective behaviours: exploring population heterogeneity of human

mobility based on social media data. *EPJ Data Science* **8** (1), p. 34. DOI: 10.1140/epjds/s13688-019-0212-x.

[11] Y. Liao, S. Yeh and J. Gil (26th Jan. 2021). Feasibility of estimating travel demand using geolocations of social media data. *Transportation*, pp. 1–25. DOI: 10.1007/s11116-021-10171-x.

[12] Y. Liao, K. Ek, E. Wennerberg, S. Yeh and J. Gil (26th Apr. 2021). *A mobility model for synthetic travel demand from sparse individual traces*. Submitted to Computers Environment and Urban Systems, Under review.

[13] Y. Liao, J. Gil, R. H. M. Pereira, S. Yeh and V. Verendel (4th Mar. 2020). Disparities in travel times between car and transit: spatiotemporal patterns in cities. *Scientific Reports* **10** (4056). DOI: 10.1038/s41598-020-61077-0.

[14] M. Schuler, B. Lepori, V. Kaufmann and D. Joye (1997). Eine integrative sicht der mobilität: im hinblick auf ein neues paradigma der mobilitätsforschung. *Bern: Schweizerischer Wissenschaftsrat*.

[15] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini and M. Tomasini (2018). Human mobility: models and applications. *Physics Reports* **734**, pp. 1–74.

[16] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil and C. Mascolo (2012). A tale of many cities: universal patterns in human urban mobility. *PloS one* **7** (5), e37027. DOI: 10.1371/journal.pone.0037027.

[17] M. Treiber and A. Kesting (2013). Traffic flow dynamics. *Traffic Flow Dynamics: Data, Models and Simulation, Springer-Verlag Berlin Heidelberg*. DOI: 10.1007/978-3-642-32460-4.

[18] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco and A. Vespignani (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences* **106** (51), pp. 21484–21489. DOI: 10.1073/pnas.0906910106.

[19] V. Kaufmann, M. M. Bergman and D. Joye (2004). Motility: mobility as capital. *International journal of urban and regional research* **28** (4), pp. 745–756.

[20] D. A. Hensher, K. J. Button, K. E. Haynes and P. R. Stopher (2004). *Handbook of transport geography and spatial systems*. Emerald Group Publishing Limited.

[21] J. Cidell and D. Prytherch (2015). *Transport, mobility, and the production of urban space*. Routledge.

[22] J.-P. Rodrigue, C. Comtois and B. Slack (2013). *The geography of transport systems*. Routledge.

[23] K. Winter, O. Cats, G. Correia and B. van Arem (2018). Performance analysis and fleet requirements of automated demand-responsive transport systems as an urban public transport service. *International journal of transportation science and technology* **7** (2), pp. 151–167.

[24] S. Shaheen, A. Cohen, I. Zohdy et al. (2016). *Shared mobility: current practices and guiding principles*. Tech. rep. United States. Federal Highway Administration.

[25] A. W. Schäfer and S. Yeh (2020). A holistic analysis of passenger travel energy and greenhouse gas intensities. *Nature Sustainability*, pp. 1–4.

[26] J. Pucher (2004). Public transportation. *The Geography of Urban Transportation* **3**, pp. 199–236.

[27] D. Banister (2011). Cities, mobility and climate change. *Journal of Transport Geography* **19** (6), pp. 1538–1546.

[28] A. Rabl and A. De Nazelle (2012). Benefits of shift from car to active transport. *Transport Policy* **19** (1), pp. 121–131.

[29] O. Edenhofer (2015). *Climate change 2014: mitigation of climate change*. Vol. 3. Cambridge University Press. Chap. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change.

[30] T. F. Welch, S. R. Gehrke and A. Widita (2020). Shared-use mobility competition: a trip-level analysis of taxi, bikeshare, and transit mode choice in washington, dc. *Transportmetrica A: transport science* **16** (1), pp. 43–55.

[31] M. C. Gonzalez, C. A. Hidalgo and A.-L. Barabasi (2008). Understanding individual human mobility patterns. *Nature* **453** (7196), pp. 779–782.

[32] J. K. Laurila, D. Gatica-Perez, I. Aad, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen et al. (2012). 'The mobile data challenge: big data for mobile computing research'. In: *Pervasive computing*. EPFL-CONF-192489.

[33] L. Alessandretti, U. Aslak and S. Lehmann (2020). The scales of human mobility. *Nature* **587** (7834), pp. 402–407.

[34] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow and C. O. Buckee (2012). Quantifying the impact of human mobility on malaria. *Science* **338** (6104), pp. 267–270.

[35] Y. Liao and S. Yeh (10th Dec. 2018). 'Predictability in human mobility based on geographical-boundary-free and long-time social media

data'. In: *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, pp. 2068–2073. DOI: 10.1109/ITSC.2018.8569770.

[36] S. Phithakkitnukoon, Z. Smoreda and P. Olivier (2012). Socio-geography of human mobility: a study using longitudinal mobile phone data. *PloS one* **7** (6), e39253.

[37] J. H. Lee, A. W. Davis, S. Y. Yoon and K. G. Goulias (2016). Activity space estimation with longitudinal observations of social media data. *Transportation* **43** (6), pp. 955–977.

[38] F. Pianese, X. An, F. Kawsar and H. Ishizuka (2013). 'Discovering and predicting user routines by differential analysis of social network traces'. In: *World of wireless, mobile and multimedia networks (wowmom), 2013 ieee 14th international symposium and workshops on a*. IEEE, pp. 1–9.

[39] T. M. T. Do, O. Dousse, M. Miettinen and D. Gatica-Perez (2015). A probabilistic kernel method for human mobility prediction with smartphones. *Pervasive and Mobile Computing* **20**, pp. 13–28.

[40] P. Jin, M. Cebelak, F. Yang, J. Zhang, C. Walton and B. Ran (2014). Location-based social networking data: exploration into use of doubly constrained gravity model for origin-destination estimation. *Transportation Research Record: Journal of the Transportation Research Board* (2430), pp. 72–82.

[41] L. Alessandretti, P. Sapiezynski, V. Sekara, S. Lehmann and A. Baronchelli (2018). Evidence for a conserved quantity in human mobility. *Nature Human Behaviour*, p. 1.

[42] C. Chen, J. Ma, Y. Susilo, Y. Liu and M. Wang (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C: emerging technologies* **68**, pp. 285–299.

[43] J. Pucher, R. Buehler, D. Merom and A. Bauman (2011). Walking and cycling in the united states, 2001–2009: evidence from the national household travel surveys. *American journal of public health* **101** (S1), S310–S317.

[44] X. Liang, J. Zhao, L. Dong and K. Xu (2013). Unraveling the origin of exponential law in intra-urban human mobility. *Scientific reports* **3**, p. 2983.

[45] M. Janzen, K. Müller and K. W. Axhausen (2017). 'Population synthesis for long-distance travel de-mand simulations using mobile phone data'. In: *6th symposium of the european association for research in transportation (heart 2017)*.

[46] Z. Wang, S. Y. He and Y. Leung (2018). Applying mobile phone data to travel behaviour research: a literature review. *Travel Behaviour and Society* **11**, pp. 141–155.

[47] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen and V. D. Blondel (2013). Unique in the crowd: the privacy bounds of human mobility. *Scientific reports* **3**, p. 1376.

[48] S. Gao, Y. Liu, Y. Wang and X. Ma (2013). Discovering spatial interaction communities from mobile phone d ata. *Transactions in GIS* **17** (3), pp. 463–481.

[49] M. S. Iqbal, C. F. Choudhury, P. Wang and M. C. González (2014). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies* **40**, pp. 63–74.

[50] G. Chen, S. Hoteit, A. C. Viana, M. Fiore and C. Sarraute (2018). Enriching sparse mobility information in call detail records. *Computer Communications* **122**, pp. 44–58.

[51] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim and S. Chong (2011). On the levy-walk nature of human mobility. *IEEE/ACM transactions on networking (TON)* **19** (3), pp. 630–643.

[52] M. De Domenico, A. Lima and M. Musolesi (2013). Interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing* **9** (6), pp. 798–807.

[53] A. Sadilek and J. Krumm (2012). 'Far out: predicting long-term human mobility.' In: *Twenty-sixth aaai conference on artificial intelligence*.

[54] V. Etter, M. Kafsi and E. Kazemi (2012). 'Been there, done that: what your mobility traces reveal about your behavior'. In: *Mobile data challenge by nokia workshop, in conjunction with int. conf. on pervasive computing*. EPFL-CONF-178426.

[55] Y. Zheng, Q. Li, Y. Chen, X. Xie and W.-Y. Ma (2008). 'Understanding mobility based on gps data'. In: *Proceedings of the 10th international conference on ubiquitous computing*. ACM, pp. 312–321.

[56] C. D. Cottrill, F. C. Pereira, F. Zhao, I. F. Dias, H. B. Lim, M. E. Ben-Akiva and P. C. Zegras (2013). Future mobility survey: experience in developing a smartphone-based travel survey in singapore. *Transportation Research Record* **2354** (1), pp. 59–67.

[57] F. Morstatter, J. Pfeffer, H. Liu and K. M. Carley (2013). 'Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose.' In: *Icwsm*.

[58] G. Stolf Jeuken (2017). 'Using big data for human mobility patterns – examining how twitter data can be used in the study of human movement across space'. MA thesis. URL: http://studentarbeten.chalmers.se/publication/250155-using-big-data-for-human-mobility-patterns-examining-how-twitter-data-can-be-used-in-the-study-of-hu.

[59] M. Lenormand, M. Picornell, O. G. Cantú-Ros, A. Tugores, T. Louail, R. Herranz, M. Barthelemy, E. Frias-Martinez and J. J. Ramasco (2014). Cross-checking different sources of mobility information. *PLoS One* **9** (8), e105184.

[60] R. Jurdak, K. Zhao, J. Liu, M. AbouJaoude, M. Cameron and D. Newth (2015). Understanding human mobility from twitter. *PloS one* **10** (7), e0131469.

[61] S. Gao, J.-A. Yang, B. Yan, Y. Hu, K. Janowicz and G. McKenzie (2014). 'Detecting origin-destination mobility flows from geotagged tweets in greater los angeles area'. In: *Eighth international conference on geographic information science (giscience'14)*. Citeseer.

[62] M. Lenormand, B. Gonçalves, A. Tugores and J. J. Ramasco (2015). Human diffusion and city influence. *Journal of The Royal Society Interface* **12** (109), p. 20150473.

[63] M. M. Hasnat and S. Hasan (2018). Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. *Transportation Research Part C: Emerging Technologies* **96**, pp. 38–54.

[64] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow and C. O. Buckee (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of the Royal Society Interface* **10** (81), p. 20120986.

[65] D. Tasse, Z. Liu, A. Sciuto and J. I. Hong (2017). 'State of the geotags: motivations and recent changes.' In: *Icwsm*, pp. 250–259.

[66] Z. Zhang, Q. He and S. Zhu (2017). Potentials of using social media to infer the longitudinal travel behavior: a sequential model-based clustering method. *Transportation Research Part C: Emerging Technologies* **85**, pp. 396–414.

[67] A. I. J. T. Ribeiro, T. H. Silva, F. Duarte-Figueiredo and A. A. Loureiro (2014). 'Studying traffic conditions by analyzing foursquare and instagram data'. In: *Proceedings of the 11th acm symposium on performance evaluation of wireless ad hoc, sensor, & ubiquitous networks*. ACM, pp. 17–24.

[68] J. H. Lee, S. Gao and K. G. Goulias (2015). 'Can twitter data be used to validate travel demand models'. In: *14th international conference on travel behaviour research*.

[69] P. A. Burrough, R. McDonnell, R. A. McDonnell and C. D. Lloyd (2015). *Principles of geographical information systems*. Oxford university press.

[70] G. R. Calegari, I. Celino and D. Peroni (2016). City data dating: emerging affinities between diverse urban datasets. *Information Systems* **57**, pp. 223–240.

[71] X. Luo, L. Dong, Y. Dou, N. Zhang, J. Ren, Y. Li, L. Sun and S. Yao (2017). Analysis on spatial-temporal features of taxis' emissions from big data informed travel patterns: a case of shanghai, china. *Journal of Cleaner Production* **142**, pp. 926–935.

[72] *OpenStreetMap* (2019). URL: `https://www.openstreetmap.org/` (Retrieved: 2019-11-13).

[73] *Google Transit APIs* (2019). URL: `https://developers.google.com/transit/` (Retrieved: 2019-11-13).

[74] *HERE Traffic* (2019). URL: `https://www.here.com/` (Retrieved: 2019-11-13).

[75] H. Ian (2010). *An introduction to geographical information systems*. Pearson Education India.

[76] Geofabrik GmbH and OpenStreetMap Contributors (2018). *OpenStreetMap Data Extracts*. URL: `http://download.geofabrik.de/` (Retrieved: 2019-11-13).

[77] G. Boeing (2017). Osmnx: new methods for acquiring, constructing, analyzing, and visualizing complex street networks. *Computers, Environment and Urban Systems* **65**, pp. 126–139.

[78] *GTFS static dataset* (2019). URL: `https://gtfs.org/reference/static` (Retrieved: 2019-11-13).

[79] *OpenMobilityData* (2019). URL: `http://transitfeeds.com/` (Retrieved: 2019-11-13).

[80] G. Lyons (2018). Getting smart about urban mobility–aligning the paradigms of smart and sustainable. *Transportation Research Part A: Policy and Practice* **115**, pp. 4–14.

[81] V. Verendel and S. Yeh (2019). Measuring traffic in cities through a large-scale online platform (in press). *Journal of Big Data Analytics in Transportation*.

[82] City of New York (May 2020). *TLC trip record data*. URL: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[83] X. Qian and S. V. Ukkusuri (2015). Spatial variation of the urban taxi ridership using gps data. *Applied Geography* **59**, pp. 31–42.

[84] C. Kamga, M. A. Yazici and A. Singhal (2015). Analysis of taxi demand and supply in new york city: implications of recent taxi regulations. *Transportation Planning and Technology* **38** (6), pp. 601–625.

[85] H. H. Hochmair (2016). Spatiotemporal pattern analysis of taxi trips in new york city. *Transportation research record* **2542** (1), pp. 45–56.

[86] F. Wang and C. L. Ross (2019). New potential for multimodal connection: exploring the relationship between taxi and transit in new york city (nyc). *Transportation* **46** (3), pp. 1051–1072.

[87] A. Gandomi and M. Haider (2015). Beyond the hype: big data concepts, methods, and analytics. *International Journal of Information Management* **35** (2), pp. 137–144.

[88] K. Crawford et al. (2011). Six provocations for big data.

[89] D. Cielen, A. Meysman and M. Ali (2016). *Introducing data science: big data, machine learning, and more, using python tools*. Manning Publications Co.

[90] E. Toch, B. Lerner, E. Ben-Zion and I. Ben-Gal (2019). Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems* **58** (3), pp. 501–523.

[91] M. Kantardzic (2011). *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons.

[92] P. J. Rousseeuw (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**, pp. 53–65.

[93] M. M. Deza and E. Deza (2009). 'Encyclopedia of distances'. In: *Encyclopedia of distances*. Springer, pp. 1–583.

[94] J. H. Ward Jr (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* **58** (301), pp. 236–244.

[95] A. K. Jain (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters* **31** (8), pp. 651–666.

[96] C. Molnar (2020). *Interpretable machine learning*. Lulu.com.

[97] T. J. Hastie and R. J. Tibshirani (1990). *Generalized additive models*. Vol. 43. CRC press.

[98] H. Nori, S. Jenkins, P. Koch and R. Caruana (2019). Interpretml: a unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*.

[99] Y. Lou, R. Caruana, J. Gehrke and G. Hooker (2013). 'Accurate intelligible models with pairwise interactions'. In: *Proceedings of the 19th acm sigkdd international conference on knowledge discovery and data mining*, pp. 623–631.

[100] K. D'Silva, A. Noulas, M. Musolesi, C. Mascolo and M. Sklar (2018). Predicting the temporal activity patterns of new venues. *EPJ data science* **7** (1), p. 13.

[101] A.-L. Barabási et al. (2016). *Network science*. Cambridge: Cambridge university press.

[102] D. Brockmann, L. Hufnagel and T. Geisel (2006). The scaling laws of human travel. *Nature* **439** (7075), pp. 462–465.

[103] G. Chen, A. C. Viana, M. Fiore and C. Sarraute (2019). Complete trajectory reconstruction from sparse mobile phone data. *EPJ Data Science* **8** (1), p. 30.

[104] X.-P. Han, Q. Hao, B.-H. Wang and T. Zhou (2011). Origin of the scaling law in human mobility: hierarchy of traffic systems. *Physical Review E* **83** (3), p. 036117.

[105] P. Plötz, N. Jakobsson and F. Sprei (2017). On the distribution of individual daily driving distances. *Transportation research part B: methodological* **101**, pp. 213–227.

[106] Z. Kou and H. Cai (2019). Understanding bike sharing travel patterns: an analysis of trip data from eight cities. *Physica A: Statistical Mechanics and its Applications* **515**, pp. 785–797.

[107] C. Anda, A. Erath and P. J. Fourie (2017). Transport modelling in the age of big data. *International Journal of Urban Sciences* **21** (sup1), pp. 19–42.

[108] A. Horni, K. Nagel and K. W. Axhausen (2016). *The multi-agent transport simulation matsim*. Ubiquity Press London.

[109] H. J. Miller (2016). Time geography and space–time prism. *International encyclopedia of geography: People, the earth, environment and technology*, pp. 1–19.

[110] S. Gambs, M.-O. Killijian and M. N. del Prado Cortez (2012). 'Next place prediction using mobility markov chains'. In: *Proceedings of the first workshop on measurement, privacy, and mobility*. ACM, p. 3.

[111] J. Petzold, F. Bagci, W. Trumler and T. Ungerer (2006). 'Comparison of different methods for next location prediction'. In: *European conference on parallel processing*. Springer, pp. 909–918.

[112] M. G. McNally (2000). The four step model.

[113] A. Peterson (2007). 'The origin-destination matrix estimation problem: analysis and computations'. PhD thesis. Institutionen för teknik och naturvetenskap.

[114] G. K. Zipf (1946). The p 1 p 2/d hypothesis: on the intercity movement of persons. *American sociological review* **11** (6), pp. 677–686.

[115] F. Yang, P. J. Jin, Y. Cheng, J. Zhang and B. Ran (2015). Origin-destination estimation for non-commuting trips using location-based social networking data. *International Journal of Sustainable Transportation* **9** (8), pp. 551–564.

[116] M. Ben-Akiva, P. P. Macke and P. S. Hsu (1985). *Alternative methods to estimate route-level trip tables and expand on-board surveys*. 1037.

[117] M. R. McCord, R. G. Mishalani, P. Goel and B. Strohl (2010). Iterative proportional fitting procedure to determine bus route passenger origin–destination flows. *Transportation Research Record* **2145** (1), pp. 59–65.

[118] S. A. Stouffer (1960). Intervening opportunities and competing migrants. *Journal of regional science* **2** (1), pp. 1–26.

[119] F. Simini, M. C. González, A. Maritan and A.-L. Barabási (2012). A universal model for mobility and migration patterns. *Nature* **484** (7392), p. 96.

[120] C. Jin, A. Nara, J.-A. Yang and M.-H. Tsou (2019). Similarity measurement on human mobility data with spatially weighted structural similarity index (spssim). *Transactions in GIS*.

[121] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli et al. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13** (4), pp. 600–612.

[122] T. Djukic, S. Hoogendoorn and H. Van Lint (2013). *Reliability assessment of dynamic od estimation methods based on structural similarity index*. Tech. rep.

[123] T. Pollard, N. Taylor, T. van Vuren and M. MacDonald (2013). 'Comparing the quality of od matrices in time and between data sources'. In: *Proceedings of the european transport conference*.

[124] T. Djukic (2014). Dynamic od demand estimation and prediction for dynamic traffic management.

[125] A. Amini, K. Kung, C. Kang, S. Sobolevsky and C. Ratti (2014). The impact of social segregation on human mobility in developing and industrialized regions. *EPJ Data Science* **3** (1), p. 6.

[126] L. Huang, Y. Yang, H. Gao, X. Zhao and Z. Du (2018). Comparing community detection algorithms in transport networks via points of interest. *IEEE Access* **6**, pp. 29729–29738.

[127] S. Sobolevsky, R. Campari, A. Belyi and C. Ratti (2014). General optimization technique for high-quality community detection in complex networks. *Physical Review E* **90** (1), p. 012811.

[128] G. Csardi and T. Nepusz (2006). The igraph software package for complex network research. *InterJournal* **Complex Systems**, p. 1695. URL: `http://igraph.org`.

[129] OpenTripPlanner developers group (2019). *Opentripplanner*. `https://github.com/opentripplanner/OpenTripPlanner`. Version 1.3.0.

[130] T. Liebig, N. Piatkowski, C. Bockermann and K. Morik (2017). Dynamic route planning with real-time traffic predictions. *Information Systems* **64**, pp. 258–265.

[131] A. F. Stewart (2017). Mapping transit accessibility: possibilities for public participation. *Transportation Research Part A: Policy and Practice* **104**, pp. 150–166.

[132] R. H. Pereira (2019). Future accessibility impacts of transport policy scenarios: equity and sensitivity to travel time thresholds for bus rapid transit expansion in rio de janeiro. *Journal of Transport Geography* **74**, pp. 321–332.

[133] M. Stępniak, J. P. Pritchard, K. T. Geurs and S. Goliszek (2019). The impact of temporal resolution on public transport accessibility measurement: review and case study in poland. *Journal of transport geography* **75**, pp. 8–24.

[134] Baidu Maps (Nov. 2020). *Transit API*. URL: `https://lbsyun.baidu.com/index.php?title=webapi/direction-api-v2`.

[135] A. Getis and J. K. Ord (2010). 'The analysis of spatial association by use of distance statistics'. In: *Perspectives on spatial data analysis*. Springer, pp. 127–145.

[136] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela and J. N. Rosenquist (2011). 'Understanding the demographics of twitter users'. In: *Fifth international aaai conference on weblogs and social media*, pp. 554–557. URL: `https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816/3234`.

[137] Y. Liao (11th May 2021). *Ride-sourcing compared to its public-transit alternative using big trip data*. Submitted to Journal of Transport Geography, Under review.

[138] J. L. Renne (2016). *Transit oriented development: making it happen*. Routledge.

[139] B. Moya-Gómez and K. T. Geurs (2018). The spatial–temporal dynamics in job accessibility by car in the netherlands during the crisis. *Regional studies*, pp. 1–12.

[140] Shared-use Mobility Center (May 2020). *What is shared mobility*. URL: `https://sharedusemobilitycenter.org/what-is-shared-mobility/`.

[141] J. Narayan, O. Cats, N. van Oort and S. Hoogendoorn (2019). 'Does ride-sourcing absorb the demand for car and public transport in amsterdam?' In: *2019 6th international conference on models and technologies for intelligent transportation systems (mt-its)*. IEEE, pp. 1–7.

[142] L. Redman, M. Friman, T. Gärling and T. Hartig (2013). Quality attributes of public transport that attract car users: a research review. *Transport policy* **25**, pp. 119–127.

[143] S. V. Walle and T. Steenberghen (2006). Space and time related determinants of public transport use in trip chains. *Transportation Research Part A: Policy and Practice* **40** (2), pp. 151–162.

[144] M. Zhou, D. Wang, Q. Li, Y. Yue, W. Tu and R. Cao (2017). Impacts of weather on public transport ridership: results from mining data from different sources. *Transportation research part C: emerging technologies* **75**, pp. 17–29.

[145] M. Zhang (2004). The role of land use in travel mode choice: evidence from boston and hong kong. *Journal of the American planning association* **70** (3), pp. 344–360.

[146] J. M. Barajas and A. Brown (2021). Not minding the gap: does ride-hailing serve transit deserts? *Journal of Transport Geography Volume 90, January 2021,* **90** (102918), pp. 1–14.

[147] F. Alemi, G. Circella, S. Handy and P. Mokhtarian (2018). What influences travelers to use uber? exploring the factors affecting the adoption of on-demand ride services in california. *Travel Behaviour and Society* **13**, pp. 88–104.

[148] A. Schafer and D. Victor (1997). The past and future of global mobility. *Scientific American* **277** (4), pp. 58–61.

[149] C. M. Schneider, V. Belik, T. Couronné, Z. Smoreda and M. C. González (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface* **10** (84), p. 20130246.

[150] G. Beirão and J. S. Cabral (2007). Understanding attitudes towards public transport and private car: a qualitative study. *Transport policy* **14** (6), pp. 478–489.

[151] N. van Oort, T. Brands and E. de Romph (2015). Short-term prediction of ridership on public transport with smart card data. *Transportation research record* **2535** (1), pp. 105–111.

[152] A. Alsger, B. Assemi, M. Mesbah and L. Ferreira (2016). Validating and improving public transport origin–destination estimation algorithm using smart card fare data. *Transportation Research Part C: Emerging Technologies* **68**, pp. 490–506.

[153] X. Zhang, Y. Xu, W. Tu and C. Ratti (2018). Do different datasets tell the same story about urban mobility—a comparative study of public transit and taxi usage. *Journal of Transport Geography* **70**, pp. 78–90.

[154] Z. Wang, S. Hale, D. I. Adelani, P. Grabowicz, T. Hartman, F. Flöck and D. Jurgens (2019). 'Demographic inference and representative population estimates from multilingual social media data'. In: *The world wide web conference*, pp. 2056–2067.

[155] E. G. Ravenstein (1885). The laws of migration. *Journal of the statistical society of London* **48** (2), pp. 167–235.