THESIS FOR THE DEGREE OF LICENTIATE OF ENGINEERING

Understanding Human Mobility with Emerging Data Sources:

Validation, spatiotemporal patterns, and transport modal disparity

YUAN LIAO

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

Göteborg, Sweden 2020

Understanding Human Mobility with Emerging Data Sources:
Validation, spatiotemporal patterns, and transport modal disparity
YUAN LIAO

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 2020

Abstract

Human mobility refers to the geographic displacement of human beings, seen as individuals or groups, in space and time. The understanding of mobility has broad relevance, e.g., how fast epidemics spread globally. After 2030, transport is likely to become the sector with the highest emissions in the 2°C scenario. Better informed policymaking requires up-to-date empirical mobility data with good quality. However, the conventional methods are limited when dealing with new challenges. The prevalence of digital technologies enables a large-scale collection of human mobility traces, through social media data and GPS-enabled devices etc, which contribute significantly to the understanding of human mobility. However, their potentials for the further application are not fully exploited.

This thesis uses emerging data sources, particularly Twitter data, to enhance the understanding of mobility and apply the obtained knowledge in the field of transport. The thesis answers three questions: Is Twitter a feasible data source to represent individual and population mobility? How are Twitter data used to reveal the spatiotemporal dynamics of mobility? How do Twitter data contribute to depicting the modal disparity of travel time by car vs public transit? In answering these questions, the methodological contribution of this thesis lies in the applied side of data science.

Using geotagged Twitter data, mobility is firstly described by abstract metrics and physical models; in Paper A to reveal the population heterogeneity of mobility patterns using data mining techniques; and in Paper B to estimate travel demand with a novel approach to address the sparsity issue of Twitter data. In Paper C, GIS techniques are applied to combine the travel demand as revealed by Twitter data and the transportation network to give a more realistic picture of the modal disparity in travel time between car and public transit in four cities in different countries at a high spatial and temporal granularity. The validation of using Twitter data in mobility study contributes to better utilisation of this low-cost mobility data source. Compared with a static picture obtained by conventional data sources, the dynamics introduced by social media data among others contribute to better-informed policymaking and transport planning.

Keywords: social media data, Twitter, mobility, travel time, travel mode, data mining, gravity model, geographical information systems

Appended publications

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

Paper A 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 B Y. Liao, S. Yeh and J. Gil (4th Mar. 2020). *Feasibility of estimating travel demand using social media data.* Working Paper.

Paper C 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`.

Author contributions

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

Paper C: YL, JG, and SY conceptualised the study. YL and VV preprocessed the data. YL, JG, and RP contributed to data processing and analysis. All authors wrote the manuscript.

ACKNOWLEDGEMENTS

I want to thank my supervisor and co-author Sonia Yeh for inspiring guidance. I remember those moments that we have the same thought and we speak it out at almost the same time. We also have moments when we disagree. However, we never hesitate to think loud and challenge each others' ideas. You are an excellent mentor, research partner, and my role model.

Thanks to Jorge Gil, my co-supervisor and co-author who joined the journey after one and a half years. We met at the beginning of my PhD study and you were the one I had always wanted to work with since then. I'm lucky to have you now. The discussion with you is always helpful and critical which drives me forward.

I also want to thank 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.

Thank you, my co-authors, Rafael H. M. Pereira, Gustavo S. Jeuken, and Vilhelm Verendel. It's been a pleasure to work with you and I learned a lot from you.

Thanks to all friends and colleagues in the division of Physical Resource Theory providing a fantastic working environment that is full of ideas. Thanks to Rasmus Einarsson, Swapnil Dhamal, Humberto Linero, Çaglar Tozluoglu, and Wasim Shoman; it is always fun and inspiring to discuss with you. Special thanks to my office mates, Xiaoming Kan, Ella Rebalski, and Ahmet Mandev who make my work full of joy.

Thanks to my family for supporting and encouraging me, especially my love, Jian, you are smart, patient, and always there for me.

Göteborg, March 2020
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 it 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].

After 2030, transport is likely to become the sector with the highest emissions in the 2°C scenario [6]. There are many ways to reduce the carbon emissions in the transport sector, 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 policy-making requires up-to-date empirical data with good quality. However, the conventional methods 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 has become feasible, through online social media platforms e.g., Twitter, GPS-enabled devices, smart card, and call detail records (CDR) etc. Unlike the data collected through household surveys, these emerging data sources are featured with the passive collection, large volume, easy 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 potentials for further application are not fully exploited.

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

Scope and contributions

Using emerging data sources, particularly Twitter data, the scope of this thesis reflects the natural process from understanding mobility to apply the obtained knowledge. The thesis answers the below questions:

- **Validation**. Is Twitter a feasible data source to represent individual and population mobility?
- **Spatiotemporal patterns**. How are Twitter data used to reveal the spatiotemporal dynamics of mobility?
- **Transport modal disparity**. How do Twitter data contribute to depicting the modal disparity of travel time by car vs public transit?

Using geotagged Twitter data, mobility is firstly described by abstract metrics and physical models in Paper A [10] to reveal the population heterogeneity of mobility patterns and in Paper B [11] to estimate travel demand. And in Paper C [12], GIS techniques are applied to combine the travel demand as revealed by Twitter data and the transportation network to give a more realistic picture of the modal disparity of travel time between car and public transit in four cities in different countries.

The first aspect examined by the thesis is the validation of using Twitter data in mobility study. 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 etc. 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 A illustrates that the fundamental patterns of population heterogeneity on mobility are well preserved in Twitter data. In addition, Paper B sheds light upon a more practical direction: geotagged tweets contribute to a reasonably good travel demand estimation with stability over time. The validation of using Twitter data in mobility contributes to better utilisation of this low-cost mobility data source.

Another aspect of this thesis is the dynamics brought by using Twitter data which naturally contain where and when people do various activities, i.e., the spatiotemporal patterns. The stream of Twitter data continuously depicts the "heartbeat" of city and the individuals' activities. These dynamics help to create a more vivid picture of mobility at both individual- and population-level compared to traditional data sources. This time-varying density map of human activities represents the attractions of places when modelling travel demand in Paper B. These dynamics help to reach a more realistic estimation of the modal disparity in travel time by car and transit in Paper C. Compared with a static picture, the dynamics contribute to better informed policy-making and transport planning.

The methodological contribution of this thesis lies in the applied side of data science with a specific focus on mobility in physics and transport. The application of data mining techniques provides new insights into the population heterogeneity of mobility (Paper A). Paper B proposes an alternative way of using geotagged tweets to tackle the sparsity issue of Twitter data. A data fusion framework in GIS is proposed in Paper C incorporating emerging data sources where Twitter data work as a proxy for time-varying travel demand. The usefulness of the framework is that it can reveal the modal disparity of travel time at a high spatial and temporal granularity.

Disposition of this thesis

This thesis is organised around a specific data source, Twitter, that develops into a series of concrete research topics/questions. The thesis consists of five chapters providing brief introduction to my research, followed by three appended papers. Chapter 2 gives further background on human mobility: how is it defined, what are the emerging data sources that deepen our understanding of it, and how is it facilitated by the transportation systems? Chapter 3 positions my 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 summarises and discusses the three appended research papers. Chapter 5 ends with general reflections on my research so far and an outlook into the future directions of further using emerging data sources in mobility for real-world applications.

CHAPTER 2

Background

This chapter first gives an overview of human mobility on its definition, scope, and applications (Section 2.1). To better understand mobility, empirical data have been widely applied. Therefore in the second section of this chapter (Section 2.2), the pros and cons of emerging data sources are introduced as compared with conventional data sources. At last, Section 2.3 reviews different modes provided by the transport system, and the emerging data sources that enhance the understanding of different modes' performance, particularly travel time in this thesis.

2.1 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 [13]: (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 (OD matrix) can be constructed with the origins and the destinations of all trips. Using this taxonomy, this chapter continues to review the literature with these two perspectives: **individual trajectories** and **networks of places**, here a "places network."

In the study of human mobility, quantitative theory seeks to answer relevant questions [14]. 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 [15], transport management [16], epidemiology [17], ecology, and social science [18].

2.2 Data sources to understand human mobility

In the last decade, the emerging data sources have significantly improved our understanding of mobility [8, 14, 19]. Common emerging data sources are call detail records (CDR), tracking apps on smart phones, GPS-enabled devices, and geotagged social media.

These data sources in human mobility 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., 20], CDR [21], and Twitter users' geotagged activity trajectories [e.g., 22]. Longitudinal datasets are often applied to reveal the patterns of individual mobility, e.g., the socio-geography of mobility [23] and the activity space estimation [24]. Because it is possible to observe the individual trajectory over a long period of time, more attention has been paid to the routine mobility [25] and next-location prediction [26]. 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 [27] and behaviour patterns at the population level [28]. The difference between the aforementioned two data forms is due to the practical trade-off between the number of individuals and data collection duration.

Here four data sources are discussed in detail: household travel surveys, CDR, GPS log data, and social media data. The main characteristics of the four 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 four 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
Time duration[c]	+++	+	+++	++
Number of individuals	++	+++	+++	+
Spatial coverage	+++	++	++	+
Trajectory completion	+	+++	++	+++
Accessibility	+++	++	+	+
Cost[d]	+++	+	++	++
Spatial resolution	+++	++	++	+++
Temporal resolution	+	+++	++	+++
Socio-demographic info.	✗	✓	✗	✓
Trip info.	✗	✓	✗	✗/✓
Passive collection	✓	✗	✓	✗

2.2.1 Household travel survey

Due to the lack of longitudinal data, most previous studies used lateral data [29] among which household travel surveys were the most prevalent. Pucher et al (2011) analysed the 2001 and 2009 National Hoursehold Travel Surveys to understanding how the daily walking and biking behaviour changes at the population level [30]. Liang et al (2013) revealed the exponential law of intro-urban mobility based on a one-year of 46, 000 trips between 2017 zones within a county [31].

Travel surveys contain socio-demographic information and detailed activity records making them not easily replaceable by other emerging data sources [32]. 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 [33]. Travel surveys also fail to capture most of the long-distance trips [32].

2.2.2 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 et al. (2012) explored geo-social radius of individuals using one year of anonymised call detail records of over one million mobile phone users in a country [23]; in order to identify the privacy bounds of human mobility, De Montjoye et al. (2013) collected data from 1.5 million users of a mobile phone operator in a country for one year [34]. In addition, the application of CDR has matured for understanding the clustering structure of spatial interactions [35] and developing OD matrices [36].

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 [21]. Nevertheless, this data source is often not easy to access, and, compared with travel surveys, has the shortcomings of spatiotemporal sparsity and incomplete trajectories [37]. It is also often not available for follow-up tracking and continuous update.

2.2.3 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 [20]). 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. (2011) revealed the Levy-walk nature of human mobility based on 101 individuals' GPS traces collected in five outdoor sites over 226 days [38]. De Domenico et al.

(2013) explored the predictability of human mobility and social interactions using a dataset collected from 25 individuals over one year in a country [39]. A large amount of studies seek the good performance of individuals' future location prediction [e.g., 40].

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., 41]. 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., 42]). Overall, GPS log data provide a relatively complete and accurate picture of individual mobility trajectory, making it close to the "ground truth."

2.2.4 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%) [43]. That number varies between regions, 7.4% (George, South Africa), 1.9% (Barcelona, Spain), 1.1% (Kuwait), and 0.3% (Sweden) [44]. Despite the low proportion of geotagged tweets, these check-ins provide precise location information and have increasingly been used for understanding mobility [45, 46].

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 [46]. 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 [27]. Gao et al. (2014) 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 [47].

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 [48]. Hasnat and Hasan (2018) used geotagged tweets to identify tourists and to study the spatial patterns of their destinations [49]. Exploring urban mobility and neighbourhood isolation, Wang et al. (2018) 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 [4].

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) [46]. Moreover, it allows for long-term tracking of movements that are free of geographical boundaries [22].

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., 50, 51] and to validate against "ground truth" [e.g., 45]. 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 [52]. 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 [53] and travel-demand data [54], 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 Mobility in transport systems

To study how mobility is facilitated 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, 55]

The essence of transportation is not planes, trains, and automobiles, but rather **mobility** and **access** [p3, 56]. The interaction between travellers and environment is emphasised when studying mobility as in transport systems, when compared to the view of physics on mobility.

A typical and relevant research topic is **transport mode choice**. 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. Transport mode is a key determinant to the emissions. 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 transportation is private car and public transit (PT). These 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.

Besides increased greenhouse gas (GHG) emissions, ever-increasing car use worldwide, in especially developing countries, has many other negative environmental impacts, including traffic congestion, land-use issues such as parking, and increased air pollution. On the flip side, PT, for example, can provide a low-cost, energy-efficient, less polluting, and socially equitable travel alternative [57, 58]. 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 [59], and combat climate change [60].

Information and communication technologies (ICT) and the trend of big data have deepened our understanding of different modes in the transport systems. Such understanding contributes to the development of sustainable mobility. Rapidly emerging data sources and geographical information systems (GIS) have significantly increased the availability and the amount of data sensed in urban transport systems [61, 62], such as traffic speed data, **taxi GPS data** [62], and **PT smart card data** [63]. The availability of real-time traffic speed data enables more advanced traveller information systems for route choice and better-informed traffic planning [64]. Emerging data sources, such as **HERE Traffic** [65] with extensive coverage of cities in 83 countries to date [64], 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 estimates of travel time and congestion level [66]. Open data standards such as **General Transit Feed Specification (GTFS)** [67] and crowdsourcing initiatives such as **OpenStreetMap (OSM)** [68] provide data and numerous new opportunities [69].

CHAPTER 3

Methodology

This thesis is organised surrounding a keyword, data, in the context of understanding human mobility. 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 [70]. 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 [71], 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." Previously, we would often ask: "I have this question, what data do I need?" Suddenly, a hammer called "big data" is handed over to us and we start searching nails everywhere. Without the right methods and questions, 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, 72]. 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 part between Computer Science/IT and Mathematics & Statistics is Machine learning under which Data mining (Section 3.1) is applied in Paper A to reveal the population heterogeneity of

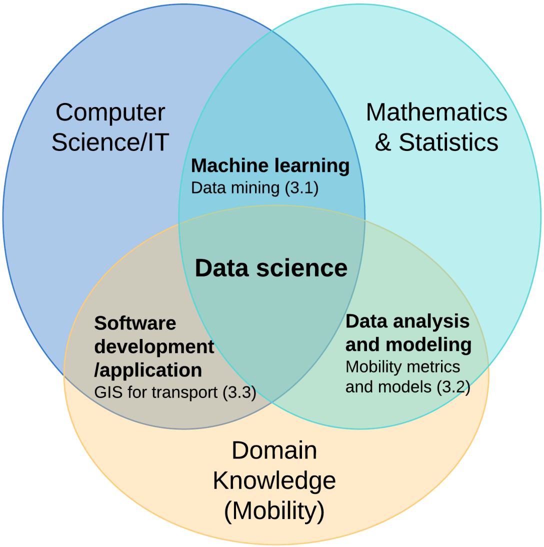

Computer Science/IT
Mathematics & Statistics
Machine learning
Data mining (3.1)
Data science
Software development /application
GIS for transport (3.3)
Data analysis and modeling
Mobility metrics and models (3.2)
Domain Knowledge (Mobility)

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

mobility. 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. In Paper C, the technology of GIS for transport (Section 3.3) applied lies in the inter-discipline of Computer Science/IT and domain knowledge of mobility; it helps to calculate the travel time of using car and taking PT in a data-driven manner. The usage of different methods are summarised in Table 3.1.

Table 3.1: Methods applied by the appended papers.

Section	Methods	Paper		
		A	B	C
3.1	Data mining	✓		
3.2	Mobility metrics and models	✓	✓	✓
3.3	GIS for transport			✓

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 [73], but choosing the right tool for a given task is vital. A detailed review can be found in the survey paper by E. Toch et al., 2019 [73].

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, 74]. There are many data-mining techniques, such as regression, classification, and clustering. Unlike some other ML techniques, such as deep learning producing 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, 74]. 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.

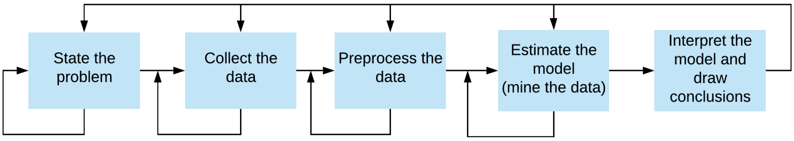

State the problem
Collect the data
Preprocess the data
Estimate the model (mine the data)
Interpret the model and draw conclusions

Figure 3.2: The data-mining process. Adapted from Figure 1.2 in [74].

The rest of this chapter introduces the particular part of data mining that has been applied in Paper A. For further information, a comprehensive description of data mining can be found in the book by M. Kantardzic, 2011 [74].

Cluster analysis: Hierarchical Clustering

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, 74]. **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 A 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. There are two categories of cluster analysis; hierarchical clustering, and iterative square-error partitional 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, 74]. 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.

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 [75], widely adopted in previous studies, is applied in Paper A. To establish cluster linkages, Ward's method was used where the decrease invariance for the cluster being merged [76]. 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 [77]. In Paper A, as

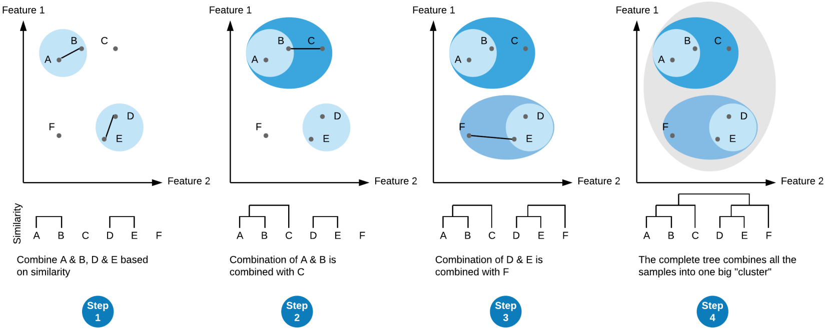

Feature 1
Feature 2
A
B
C
D
E
F
Similarity
Combine A & B, D & E based on similarity
Combination of A & B is combined with C
Combination of D & E is combined with F
The complete tree combines all the samples into one big "cluster"
Step 1
Step 2
Step 3
Step 4

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 result of cluster analysis, 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 A
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

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

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

trajectory in between (Section 2.1). The corresponding metrics and models are summarised in Figure 3.4.

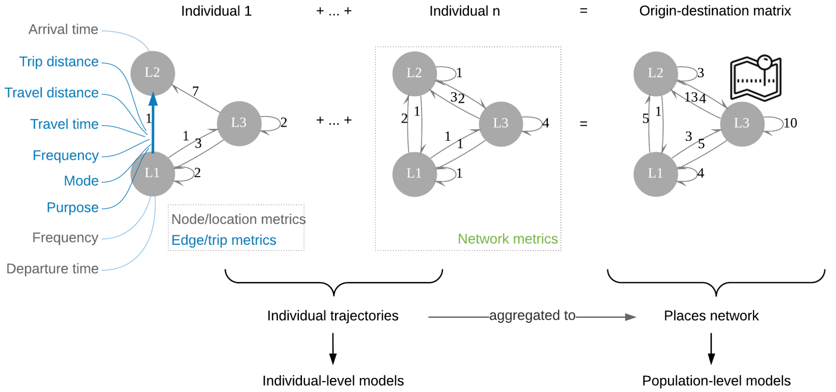


Figure 3.4: A framework of mobility metrics and models. L1-3 are three distinct locations/zones. The edges/arrows pointing from one location to another are trips that connect an origin and a destination. The numbers next to the edges are the frequency of the observed trips based on the individual trajectory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

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:

$$\mathbf{L}'_p = (X, Y, \mathbf{T})_{p,i}, i = 1, 2, \ldots, n_p \tag{3.2}$$

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 [14]. **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 ($\mathbf{G}_p$).

$$\mathbf{G}_p = (d, D, TT, m, f, P)_{i,j}, i, j = 1, 2, \ldots, n_p \tag{3.3}$$

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

$$\mathbf{G} = (d, F)_{i,j}, i, j = 1, 2, \ldots, N \tag{3.4}$$

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). From city dwellers that commute to work on a weekday morning to visitors who arrive in town for business or leisure, the urban landscape is transforming at a fast pace [78]. 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.

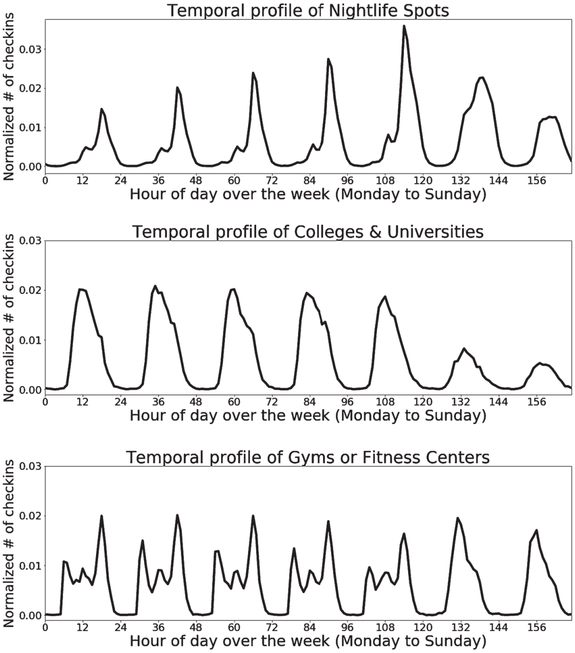


Figure 3.5: Distinct temporal profiles of different venues. Source: Figure 2 from [78].

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 [14]. 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:

$$r_g = \sqrt{\frac{1}{n_p}\sum_{i=1}^{n_p} f_i \cdot (\mathbf{r}_i - \mathbf{r}_{cm})^2} \tag{3.5}$$

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

$$\mathbf{r}_{cm} = \left[\frac{\sum_{i=1}^{n_p}(X_i \cdot f_i)}{\sum_{i=1}^{n_p} X_i}, \frac{\sum_{i=1}^{n_p}(Y_i \cdot f_i)}{\sum_{i=1}^{n_p} Y_i}\right] \tag{3.6}$$

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, 79]. For a node (location) i with degree (visiting frequency) $f_{p,i}$, its local clustering coefficient is defined as:

$$C_i = \frac{2L_i}{f_i(f_i - 1)} \tag{3.7}$$

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:

$$\overline{C} = \frac{1}{n_p}\sum_{i=1}^{n_p} C_i \tag{3.8}$$

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:

$$z = \frac{\sum_{i=1}^{n_p} \log(f_i)}{n_p} \tag{3.9}$$

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:

$$z_m = \frac{\max[f_i]}{\sum_{i=1}^{n_p} f_i} \tag{3.10}$$

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.

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 [14].

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

$$\mathbf{L}\left(t_{n_p}\right) = \sum_{i=1}^{N_p} \Delta\mathbf{L}(t_i) \tag{3.11}$$

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.

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

$$R(t) = \sqrt{\langle \mathbf{L}(t)^2 \rangle} \tag{3.12}$$

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:

$$f(\Delta L) \sim \frac{1}{\Delta L^{1+\alpha}} \tag{3.13}$$

$$\phi(\Delta t) \sim \frac{1}{\Delta L^{1+\beta}} \tag{3.14}$$

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. [14].

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., 2010 [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., 2010 [1] extended the CTRW model with the exploration and preferential return as briefly illustrated in Figure 3.6. They found:

$$\langle \Delta L^2 \rangle^{\alpha/2} \sim \log\left(\frac{1 - S^{1-\zeta}}{\zeta - 1}\right) + \text{const} \tag{3.15}$$

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.

Another stream of individual mobility models stems from Transportation and Computer Science. These models further incorporate built

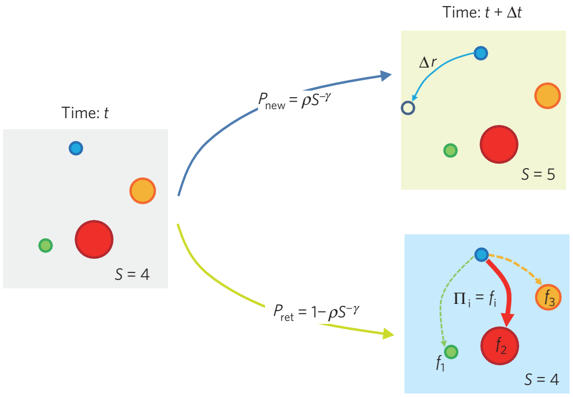


Figure 3.6: Schematic description of the individual-mobility model. Time t panel shows the starting time when historically an individual visited four locations, $S = 4$. Circles' size are proportional to their visiting frequency, f_i. For time $t + \Delta t$, this individual either visits a new location at distance Δr that follows a fat-tailed $P(\Delta r)$, or he/she returns to a previously visited location with probability $P_{ret} = 1 - \rho S^{-\gamma}$ where the next location will be chosen with probability $\Pi_i = f_i$. Source: Figure 2 from [1].

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 [80], as illustrated by the space-time prism in time geography in Figure 3.7. 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 [81].

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., 2014 applied a probabilistic kernel method for human mobility prediction with smartphones [26]. Other common methods include Markov models [83], dynamic Bayesian network, multi-layer perception, and state

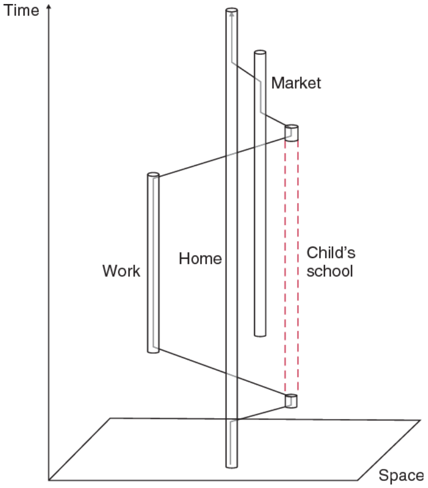


Figure 3.7: A space–time path among activity stations. Source: Figure 1 from [82].

predictor [84].

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 [14] on mobility physics, [73] on mobility models and machine learning, and [80] on big data and transport modelling.

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 [14]. 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 [85] as shown in Figure 3.8. 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 [80]. 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.

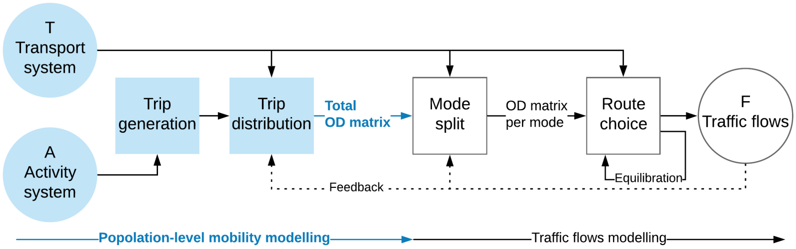


Figure 3.8: The Four Step Model. Adapted from Figure 2 in [85] and Figure 1 in [86].

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 socio-economic characteristics of the populations of i and j, and their spatial distribution. Barbosa et al., 2018 summarise a few mobility models to describe the total OD matrix [14]. 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 flows inspired by Newton's law of gravitation [87] and later on became one of the most applied methods for the trip distribution [88]. 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

$$F_{i,j} = k f_i f_j f\left(d_{i,j}\right) \tag{3.16}$$

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

$$f\left(d_{i,j}\right) = \alpha e^{-\beta d_{i,j}} \tag{3.17}$$

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) [89, 90]. 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 [91] 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., 2012 [92] 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

$$\langle F_{i,j} \rangle = f_i \frac{P_i P_j}{\left(P_i + s_{i,j}\right)\left(P_i + P_j + s_{i,j}\right)} \tag{3.18}$$

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)** [93] as used in Paper B. SpSSIM is an extended version of the original structural similarity (SSIM) proposed by [94]. 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 [95, 96]. This newly proposed SpSSIM [93] overcomes the SSIM sensitivity issue due to the ordering of OD pairs, as raised by previous studies [e.g., 97]. 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 [79, 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., 2010 [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 [98]. Using CDR datasets, the community structure detected is shown in Figure 3.9 where we can see the clear discrepancy between the administrative boundary and the naturally formulated mobility partitioning (community structure). Huang et al., 2018 compare different community detection algorithms [99] in transport networks and find the Combo algorithm [100] outperforms the other algorithms, such as the Walktrap.

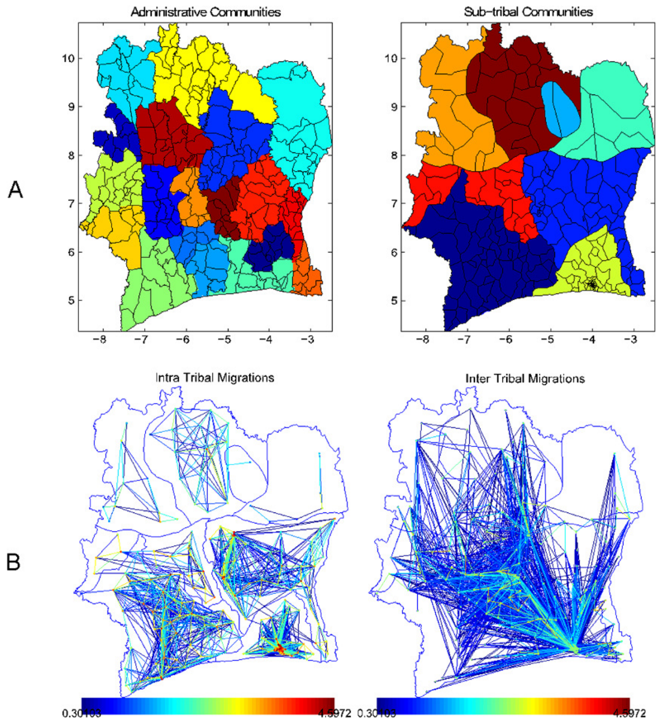


Figure 3.9: Hierarchical boundary and human migration visualisation in Ivory Coast. (A) Partitionings of Ivory Coast by administrative prefectures/sub-prefectures (left) and tribal/sub-tribal communities (right). (B) Intra-Inter tribal migrations, where each node represents an individual sub-tribal community, and each link is logarithmically coloured to represent the number of migrations (extracted from call records) between the two nodes. Source: Figure 3 in [2].

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 Geographical Information Systems (GIS) for transport

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, 101]. As shown in Figure 3.10, 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.

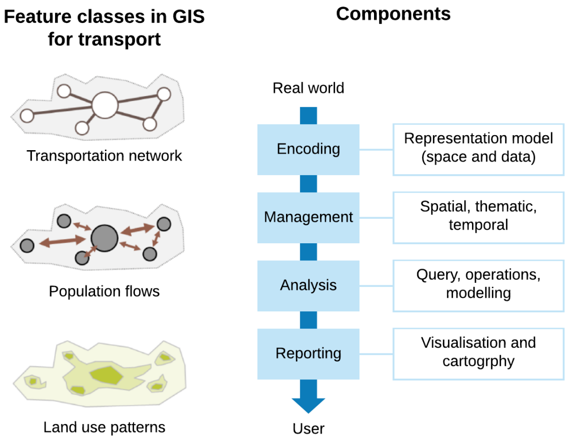

Feature classes in GIS for transport
Components
Transportation network
Population flows
Land use patterns
Real world
Encoding
Management
Analysis
Reporting
User
Representation model (space and data)
Spatial, thematic, temporal
Query, operations, modelling
Visualisation and cartogrphy

Figure 3.10: Components of GIS and major classes for transportation.[2]

Network analysis is the core function of GIS for mobility as in transport systems. 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, 102]. A network graph defines potential movements from node (place) to node including prohibited and permitted 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, 55]. Transport networks including rich attributes e.g., distance and speed limit of each network link are available via OpenStreetMap [103], a collaborative project to

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

create a free editable map of the world. An example of downloaded street network of Modena, Italy is shown in Figure 3.11 using a Python package, osmnx [104].

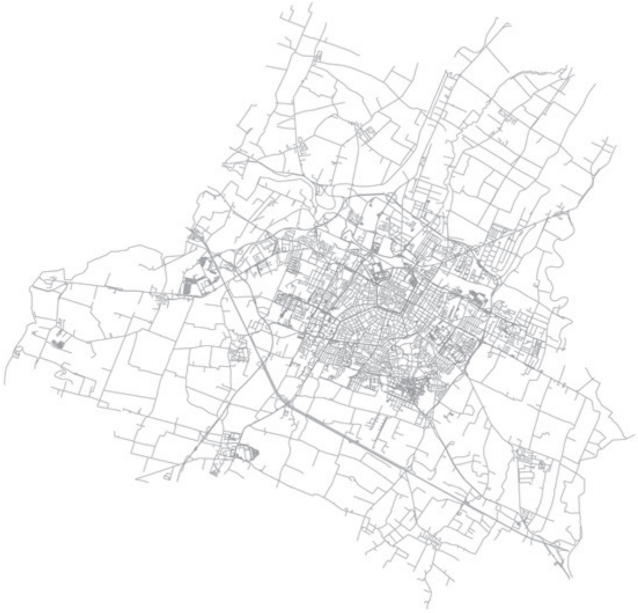


Figure 3.11: OSMnx street networks automatically downloaded and visulised for Modena, Italy. Adapted from the source: Figure 4 in [104].

Under network analysis, solving **the shortest path problem** is a key function which is particularly useful for calculating travel time ($TT_{i,j}$). Oftentimes, millions of shortest-path calculations are required to be done efficiently. To answer the question, "**How long does it take for one to drive from L1 to L2 in Stockholm considering the real traffic?**", we need a solution to finish massive routing requests in an acceptable time period and it is flexible enough to integrate the real-world measurements of road speed. As done in Paper C, the downloaded drive road network is converted into an igraph object [105] with edited links which have the hourly average speed assigned as the routing impedance based on HERE Traffic data [65].

The complexity of the shortest path problem increases as we move from calculating travel time by car to PT, because it requires intermodal 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 have been widely applied to calculate the travel time by PT. A GTFS static dataset [106] 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. Figure 3.12 shows an example of PT lines contained in a GTFS dataset from Stockholm.

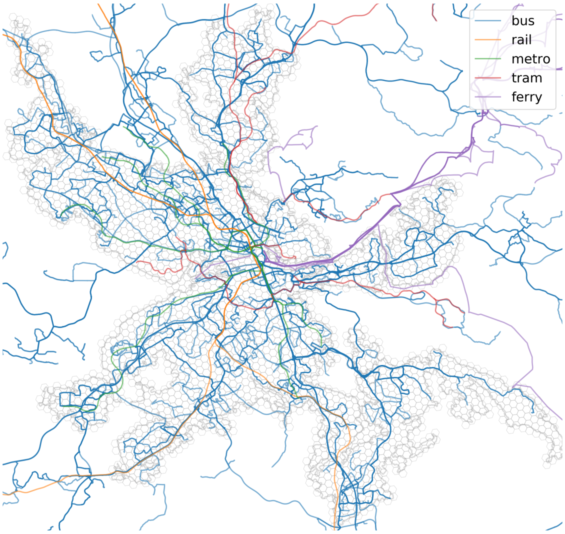


Figure 3.12: PT lines in Stockholm. Figure by Liao and Gil (ongoing study).

The actual routing process can be supported by various GIS solutions among which, OpenTripPlanner (OTP) is an open-source multimodal routing engine [107], similarly used in previous studies [108–110]. 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.

Besides network analysis, the other two feature classes, population flows and land use patterns correspond to the applications of GIS in travel demand modelling and urban planning. They are crutial aspects of studying human mobility with GIS, however, they are not within the scope of this thesis. A more comprehensive introduction is presented in [111].

CHAPTER 4

Present work

Using emerging data sources, particularly Twitter data, the three appended papers 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 questions:

- **Validation**. Is Twitter a feasible data source to represent individual and population mobility?
- **Spatiotemporal patterns**. How are Twitter data used to reveal the spatiotemporal dynamics of mobility?
- **Transport modal disparity**. How do Twitter data contribute to depicting the modal disparity of travel time by car vs public transit?

An overview of the research scope and the involved data sources are presented in Figure 4.1. Paper A [10] and Paper B [11] demonstrate how the geotagged tweets can be applied to understand human mobility; Paper A focuses on the aspect of individual trajectories to reveal the population heterogeneity on the spatiotemporal patterns of mobility while Paper B focuses on the travel demand estimation (places network) aggregating individual trajectories. Both papers validate the results from Twitter data against some established data sources to reveal the feasibility of using this emerging data source. These two papers analyse the empirical results of mobility in space and time while how people travel from one place to another is not considered. This gap between the mobility outcomes and the built environment is bridged in Paper C [12].

Paper C [12] reveals the disparities in travel time between car and PT in four cities. A combination of multiple emerging data sources

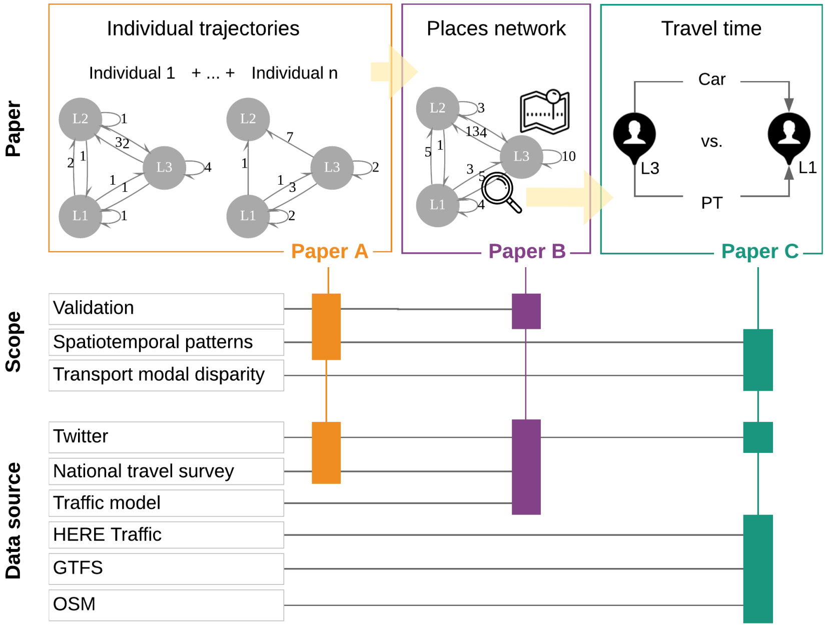


Figure 4.1: Overview of included studies: their scope and involved data sources. L1 - L3 are three distinct locations visited by a group of individuals.

empowers a finer depiction of the spatiotemporal patterns than the previous studies. The role of Twitter data is to provide the dynamics of travel demand. Therefore, Paper C can be regarded as an application of Twitter data in real-world settings.

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 A)

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 2.2.4) 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 2.2.4), includes more than 650 thousands of geotagged tweets by nearly 3 thousands of 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 2.2.4, 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 dairy.

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. 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

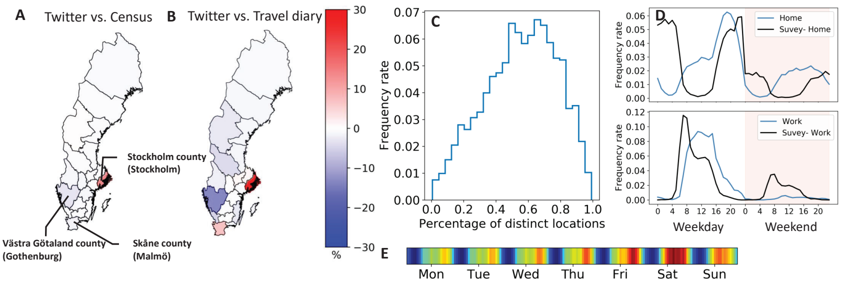


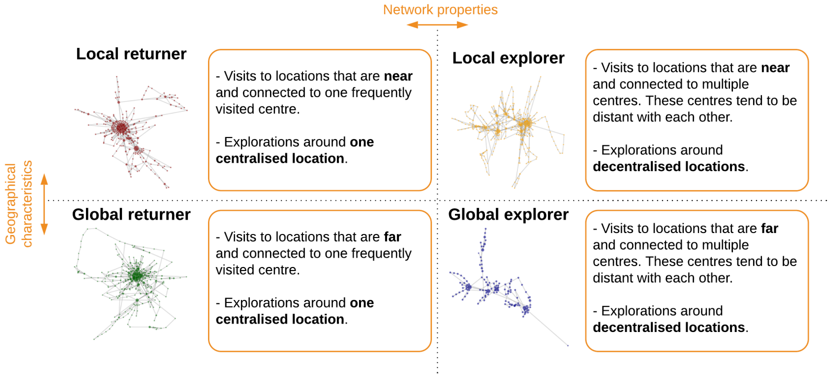
Figure 4.2: Characteristics of geotagged activity of Swedish Twitter users (adapted from Figure 2 and Figure 3 in Paper A). (A) and (B) show the county-level geographical representativeness of estimated home locations from Twitter data: percentage value difference. (**A**) Twitter users vs. residents (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.

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. However, 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**.



- Visits to locations that are **near** and connected to one frequently visited centre.

- Explorations around **one centralised location**.

- Visits to locations that are **near** and connected to multiple centres. These centres tend to be distant with each other.

- Explorations around **decentralised locations**.

- Visits to locations that are **far** and connected to one frequently visited centre.

- Explorations around **one centralised location**.

- Visits to locations that are **far** and connected to multiple centres. These centres tend to be distant with each other.

- Explorations around **decentralised locations**.

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 A). In the visualised networks, each node represents one visited location. The diameter of the node is proportional to the node degree.

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-

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
	(%)	0	1	2		(/day)
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

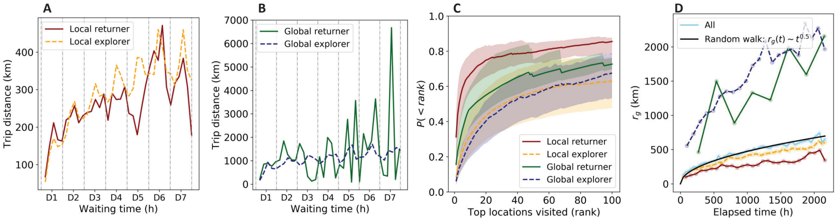


Figure 4.4: Collective mobility behaviours (adapted from Figure 9 and 10 in Paper A). Trip distance vs. waiting time during 7 days for (**A**) local travellers and (**B**) global travellers. Waiting time is defined as the time interval between two consecutive geotagged tweets generated by the same Twitter 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.

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 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 A 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 associated with these data. We validate against the data from a household 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 B)

Feasibility of Estimating Travel Demand using 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 four areas: 1) **Home/worplace locations**. The basic temporal technique to identify home/workplace is widely applied when using geotagged tweets. Our preliminary results suggest that the reliability of identifying home and workplace locations needs 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 results can be scale-dependent; 3) **Sampling methods**. Existing literature does not clearly explore how different sampling methods 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 B comprehensively examines the validity of using geotagged tweets collected from the Streaming API and User Timeline API to approximate the OD matrix at different spatial scales. We compare these Twitter-based OD matrices with the Swedish national travel survey and the traffic models' outputs from Swedish Transport Administration (Trafikverket). Specifically, we attempt to answer the below questions:

- Is Twitter a feasible data source to represent commuting travel demand?
- Can Twitter data be used to create models for travel demand estimation?
- How do spatial scale, sampling method of Twitter data, and sample size affect its representativeness for travel demand?

In order to examine the feasibility of using Twitter data for travel demand estimation, we propose a comparison framework to compare Twitter with the other established data sources. In practice, transport planners collect empirical trips from a small sample of the population and create a model to simulate the travel demand of the overall population for further application such as traffic flows modelling. Therefore, we divide the comparison work into two focuses: empirical trip records and model output.

We first compare the empirical trip records obtained from Twitter with the ones from travel survey data on the overall travel demand for an average weekday and commuting travel demand. In this part of validation, we also examine the stability of the similarity between Twitter and the travel survey over time. After the analysis of the empirical trips, we create the gravity models based on Twitter data, collected with two sampling methods, to simulate the overall travel demand at both national (long-distance travel above 100 km) and city level. In this part of validation, we use two methods for the step of trip generation followed by the gravity model for the trip distribution; they are trips converted from displacements by adding a time threshold (Model a) and the density-based approach proposed in this study (Model b). Model b is proposed as an alternative to Model a to solve the sparsity issue of Twitter data. Finally, we evaluate the comparison results by comparing the Twitter-based trips and model outcomes with the ones from the national travel survey and the Sampers model.

The comparison techniques include the visualisation of the OD matrices, the similarity measure (SpSSIM) between the OD matrix from Twitter and the external sources. An essential aspect of human mobility behaviour is the travel distance (d, km) whose distribution of the OD matrices reveals another facet of the validity of using Twitter to estimate travel demand. Therefore, we compare this distribution of Twitter data with other sources.

Main findings

Twitter for commuting travel demand estimation

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 = 0.3), and the commuting distance distribution.

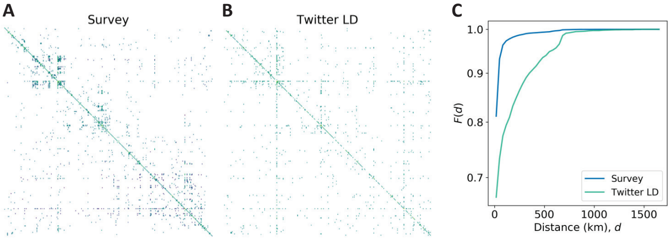


Figure 4.5: Evaluation of the feasibility of using Twitter for commuting travel demand estimation (adapted from Figure 4 and 5 in Paper B). Commuting OD matrices based on (**A**) Survey and (**B**) Twitter LD. (**C**) Commuting trip distance distribution produced by Twitter LD in comparison with Survey. $F(d)$ is defined as the probability of commuting between zones at a distance below d.

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 A. 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. Therefore, **Twitter data are not appropriate for estimating commuting travel demand**.

Demand model construction using Twitter data

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

Table 4.2: The similarity between the modelled OD matrices using Twitter data and other traffic model's outputs. * Displacements converted. Model a - displacement conversion + gravity model; Model b - density-based approach + gravity model. For all models, $\beta = 0.03$.

Scale	Model	Twitter	SpSSIM
Nation	a	LD	0.72
		LT	0.67
	b	LD	0.83
		LT	0.81
City	a	LD	0.79
		LT	0.66
	b	LD	0.87
		LT	0.88

In terms of the effect of spatial scale, Twitter data work generally well at both spatial scales (0.67 - 0.88). Comparing the two spatial scales, the greater number of traffic zones and larger geographical coverage make the national level more challenging to model using Twitter data, leading to smaller values of similarity in general than the city level, which is due to the sparsity issue. **Twitter data suit better when 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**. Using geotagged tweets for travel demand estimation requires reasonable spatial aggregation which depends on the form of Twitter data and the penetration of Twitter.

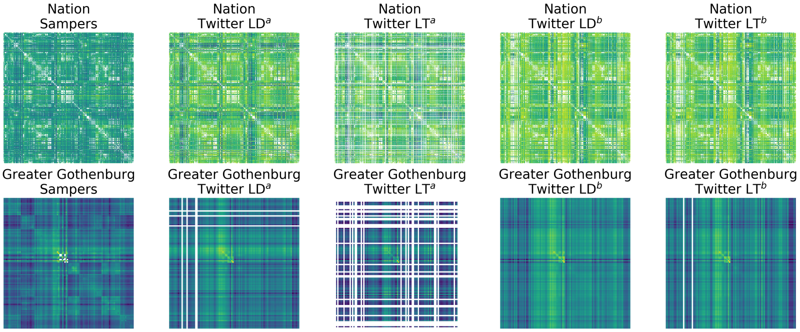


Figure 4.6: Estimated OD matrices using gravity model in the comparison with Sampers' model outputs (Figure 8 in Paper B). ^{a} Displacement conversion + gravity model. ^{b} Density-based approach + gravity model.

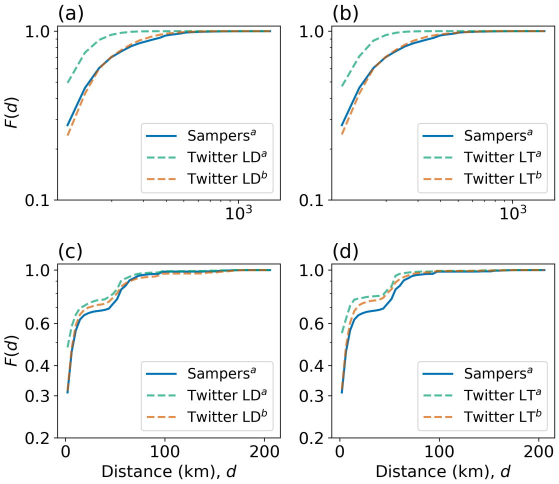


Figure 4.7: Trip distance distribution. $F(d)$ is defined as the probability of travel between zones at a distance below d. The trip distance is from the estimated OD matrices by a displacement conversion + gravity model and by b density-based approach + gravity model. (a) National level - Twitter LD. (b) National level - Twitter LT. (c) City level - Twitter LD. (d) City level - Twitter LT.

In terms of the impact of sampling method of Twitter data, **the longitudinal geotagged tweets (Twitter LD) collected from the User Timeline API have more advantages over the lateral geotagged tweets (Twitter LT) collected from the Streaming API, especially**

when the data sparsity issue is salient. Paper B illustrates that the long-term observation of longitudinal geotagged tweets by top users compensates for the time sparsity and helps to recreate a more complete image of individual mobility, and, therefore, is more reliable for the travel demand estimation than the lateral dataset despite a more than 3-fold greater sample size of the covered individuals. However, when using the density-based approach to make more geotagged tweets available, the gap between the two sampling methods is narrowed.

Another contribution of Paper B is the proposed alternative way of utilising geotagged tweets, the density-based approach. **By doing so, we bypass adding a time threshold that causes the reduction of available Twitter data for travel demand estimation**. It is inevitable to lose data when converting the geotagged displacements into geotagged trips. After adding the time threshold filter, only 17-21% of geotagged tweets are utilised to estimate the overall travel demand using the gravity model. This reduction limits the further application of geotagged tweets given the sparsity is already one of its drawbacks. Without sacrificing the similarity between Twitter data and the other data sources on the estimated OD matrices, the density-based approach generates good results while increasing the available data by 4 times more. Moreover, the density-based approach produces better trip distance distribution than the displacement conversion.

Stability of using Twitter for travel demand estimation

To examine the stability of the trips from Twitter data over time when compared with the travel survey, it is necessary to reveal the similarity of OD matrices from 2011 to 2016 at the national level. The trips aggregated each year have rather stable similarity between the travel survey as compared to the baseline year (2011), and between Twitter LD and the travel survey over time (see Figure 4.8). **The stability of this similarity between Twitter LD and the travel survey suggests good potentials of using Twitter data to estimate the national-level travel demand, especially given its low cost to continuously update**.

To further look at the sensitivity of the model outcomes to the sample size and the involved geotagged tweets, we test a share of geotagged tweets from 1% to 99%, with a step length of 1% and 10 repetitions of random sampling, to create models using Model a and

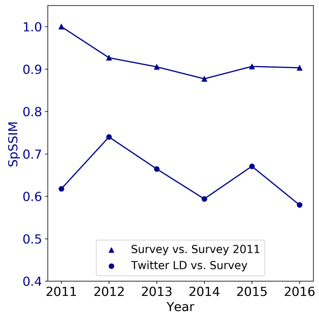


Figure 4.8: Similarity between the OD matrices from Survey and Twitter LD over time. The curve of Survey shows how the OD matrix deviates from the baseline year, 2011 for Survey records. Source: Figure 6 in Paper B.

b with the same settings as aforementioned. The similarity results are illustrated in Figure 4.9. The more geotagged tweets included in the modelling, the greater and more stable of the similarity between the Twitter-based OD matrix and the Sampers model output. However, Model a is more sensitive to the number of geotagged tweets than Model b, especially for the national level, because the number and the geographical coverage of traffic zones involved at the national level are greater than the city level, therefore it is more sensitive to the sparsity issue of Twitter data. In general, **Twitter data with the density-based approach present good stability to the sample size and the involved geotagged tweets while Twitter LD works equally or better than Twitter LT despite a much smaller number of individuals covered**.

Conclusions

Geotagged tweets are proved to be a good data source for the overall travel demand estimation for an average weekday, especially at the city level when the number of traffic zones is smaller than the national level. The high similarity of the estimated travel demand to the results based on the national travel survey remains stable year by year (2011 - 2016). However, Twitter data are not appropriate for estimating commuting travel demand due to the unreliability of the connection between the identified workplace and home. Despite a smaller group of the covered population, the longitudinal geotagged tweets collec-

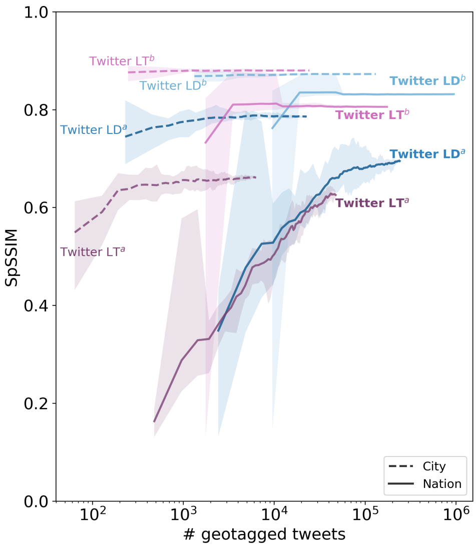


Figure 4.9: Similarity as a function of included geotagged tweets. Purple colors show the results using Twitter LT and blue colors show the results using Twitter LD. For the 10 model outcomes of each share, the curve shows the average value of SpSSIM and the shaded area shows the maximum and minimum value of SpSSIM. Model a - displacement conversion + gravity model; Model b - density-based approach + gravity model. For all models, $\beta = 0.03$.

ted from the User Timeline API have more advantages over the lateral geotagged tweets collected from the Streaming API. As for the impact of sample size, the more geotagged tweets included in the modelling, the better Twitter works for travel demand estimation. In addition, we propose a density-based approach to address the sparsity issue of geotagged tweets. When using the proposed method, the difference between the estimated travel demand of the two sampling methods is narrowed and moreover, Twitter data with the density-based approach present good stability to the sample size and the involved geotagged tweets. This density-based approach produces a better similarity than the common approach of adding an arbitrary time-threshold filter when generating trips, especially when the sparsity issue is salient. The approach increases the amount of available geotagged tweets significantly which leaves potentials for using Twitter data at a finer spatiotemporal resolution.

4.3 Modal disparity in travel time (Paper C)

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

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 [112, 113] starts using detailed spatial data and time-varying transport data sets, 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 C 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 4.10. 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.

The outcomes of the improved travel time calculations demonstrate the usefulness of applying large datasets in the framework developed in Paper C. **It is shown how the travel time for each mode changes**

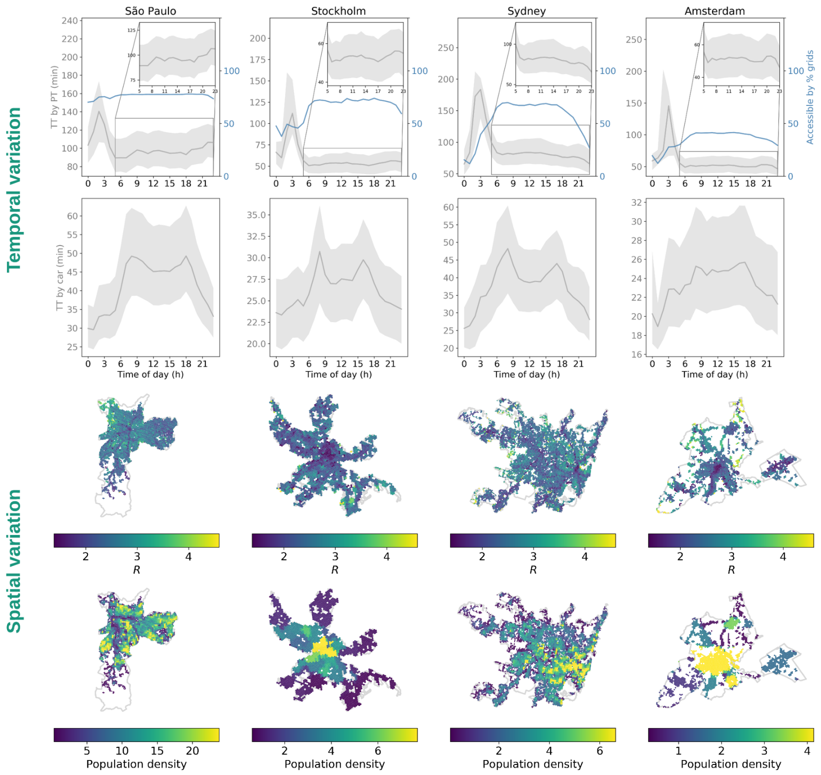


Figure 4.10: Spatiotemporal patterns of modal disparity in travel time (adapted from Figure 1 and 3 in Paper C). For Temporal variation, travel time by PT (upper row) and car (bottom row) are presented over the course of an average weekday. For Spatial variation, travel time ratio (R) to frequently visited locations (top row) and population density (bottom row) in 1000 persons per sq. km are presented.

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 4.11. **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 4.11A). **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 4.11(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 4.11(C), population density and R are also correlated: **The greater the population density, the lesser the disparity between PT and car travel times**.

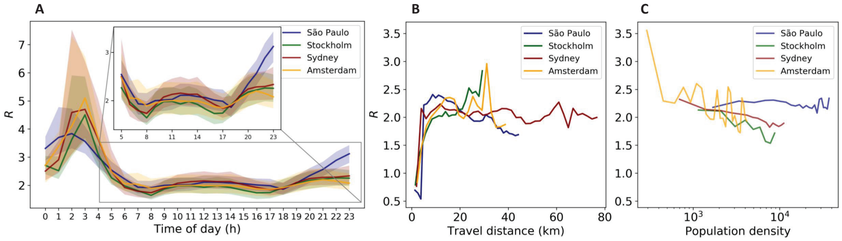


Figure 4.11: Travel time ratio across four cities (adapted from Figure 4-5 in Paper C). (**A**) Temporal variation of citywide average travel time ratio (R). The shaded area indicates the two mid quartiles. The insert zooms in on the period from 05 hours to 23 hours, to better show the temporal variation 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.

Paper C further summarises the city level performance of PT and car use in terms of the travel time ratio and the aggregate travel speed (Table 4.3). **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.

Table 4.3: 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 C 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.

More over, Paper C demonstrates 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.

CHAPTER 5

Discussion and outlook

The last decade witnessed a rapidly growing body of literature using social media data in mobility studies. The main drivers are listed below.

- 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.

This trend started with the descriptive analysis using mobility metrics and models to reproduce the observed patterns in the previous research in physics and transportation. Gradually, the research gap has been narrowed down to a more practical direction; how to use social media data in real-world settings, e.g., to guide transport planning, and what improvements we need to make to the existing methods and data itself. On the other side of this process, with deepened understanding of the pros and cons of social media data, we started formulating the research questions that can be answered by using social media data.

In this thesis, using geotagged Twitter data, mobility is firstly described by abstract metrics and physical models in Paper A to reveal the population heterogeneity of mobility patterns and in Paper B to estimate travel demand. In Paper C, GIS techniques are used to connect mobility outcomes as revealed by Twitter data and the transportation network to give a more realistic picture of the modal disparity of travel time between car and public transit in four cities in different countries.

Using emerging data sources, particularly Twitter data, the scope of this thesis reflects the natural process from understanding mobility to apply the obtained knowledge. The thesis answers the below questions:

- **Validation**. Is Twitter a feasible data source to represent individual and population mobility?

  One key strength of social media data is the low cost when compared to the traditional data sources. However, this low cost comes with a price: significant biases and incompletion. Paper A and B attempt to validate Twitter data against some external data sources to identify the potentials of this data source in representing actual mobility at individual and population level. Despite having clear signs of overly representing big-city residents and their leisure activities, mobility regularity, diffusive nature, and returning effect are preserved in the geotagged tweets to some extent. Paper A illustrates that the fundamental patterns of population heterogeneity on mobility are well preserved in Twitter data. In addition, Paper B sheds light upon a more practical direction: geotagged tweets contribute to a reasonably good travel demand estimation with stability over time. In the validation aspect of Paper A and B, a more detailed exploration is presented on the impact of Twitter data form and spatial scale etc. However, what remains a puzzle is a universal de-biasing approach that can be implemented to the data itself so it's applications can be better expanded.

- **Spatiotemporal patterns**. How are Twitter data used to reveal the spatiotemporal dynamics of mobility?

  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 "heartbeat" of city and the individuals' activities. These dynamics help to create a more vivid picture of mobility at both individual and population level. Tasse et al., 2017 [51] 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 zones in cities. In Paper B, this density map is applied to represent the attractions of places when modelling travel demand.

- **Transport modal disparity**. How do Twitter data contribute to depicting the modal disparity of travel time by car vs public transit?

  Given the importance of the spatiotemporal patterns of travel time disparity for transport planning, Paper C explores the method and validity of using Twitter data to represent time-varying demand in contrast with 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 OD matrix output from static models. Under this background, the data fusion used in Paper C is a novel approach that allows us to combine both transport service demand and operations while getting more granular results, especially through the use of Twitter data as a proxy for time-varying travel demand. And due to the easy access of geotagged tweets globally, this application can be generalised to multiple regions.

The methodological contribution of this thesis lies in the applied side of data science with a specific focus on mobility in physics and transport. The application of data mining techniques provides new insights into the population heterogeneity of mobility underlying Twitter data (Paper A). Although using a widely applied gravity model, Paper B proposes an alternative way of using geotagged tweets to tackle the sparsity issue of Twitter data. Paper C proposes a 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 in different countries. The usefulness of the framework is that it can reveal the modal disparity of travel time at a high spatial and temporal granularity.

This thesis is organised around a specific data source, Twitter data, that develops into a series of concrete research topics/questions. The risk of being "data-centric" is that we might lose the sharpness of asking the right questions and let the data lead our way. However, the "data-centric" process is necessary, especially for the application of emerging data sources in the long run. The research in the appended papers is also exploratory and slightly starting to move from the fundamental side to the application side. The further one pushes the use of social media data to the real world, the more problems one faces. However, we can make use of the obtained knowledge to tailor the use of social media data to better ask both right and relevant questions.

Outlook

The use of emerging data sources in mobility has gone through the exploratory stage, towards the application side. For the next stage of my study, there are three potential research directions to pursue. As a continuous effort following Paper A and B, the first two are about using social media data to answer relevant questions. Following the work in Paper C, the third research direction focuses on using online traffic data and more GIS data sources to evaluate the potentials of carbon emissions reduction in the transport sector of urban areas.

(1) **The characterisation of long-distance travel behaviour using social media data**.

Long-distance travel has rapidly increased in recent decades contributing to a majority of the total climate impact in the transport sector. To date, daily and short-distance trips have been extensively studied by transportation and geographic researchers using traditional household travel surveys where, however, a tendency exists to underestimate the long-distance travels of which the patterns and frequencies are often poorly characterised. Social media data collected from Twitter are especially valuable for characterising international mobility and long-distance travel as Twitter users tend to report uncommon places that are outside their daily mobility range to communicate with the followers where they've been.

(2) **Reconstruction of the individual mobility trajectories from social media data for a better estimation of travel demand**.

The picture of individual mobility obtained from geotagged tweets is incomplete due to the time sparsity and selective bias. In order to get an unbiased estimate of mobility patterns, it would be important to fill in the missing stays so that the actual weekly activity chain can be recovered, which contributes to transport planning and provides an economical way to keep the travel demand estimation up to date. This direction aims to create a model for reconstructing individual mobility trajectories from the users' geotagged tweets. The model will represent what Twitter users are doing on a weekly basis, as well as the time and place of each activity. The model will be generalised into 23 regions globally which are expected to exhibit different mobility characteristics where a cross-regional comparison of the heterogeneity of travellers will be explored.

(3) **Spatial analysis of emission reduction potentials: using the value of time to quantify the cost of reduced emissions resulting from the modal shift from high- to low-carbon intensity**.

The transport sector accounts for big share of carbon emissions. It is particularly valuable to seek for the concrete policy implications with the results to minimise the carbon footprint from the transport sector in cities. HERE Traffic data, OSM, and GTFS data provide rich information for exploring the carbon emission reduction potentials in cities at a fine spatial granularity. This direction asks "what-if" questions to look at the impact of modal shift in the transport system.

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 (4th Mar. 2020). *Feasibility of estimating travel demand using social media data.* Working Paper.

[12] 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`.

[13] 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.*

[14] 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.

[15] 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`.

[16] 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`.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] Y. Liao and S. Yeh (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.

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

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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).*

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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.

[43] 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.*

[44] 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`.

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

[53] 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.

[54] 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.*

[55] 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.

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

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

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

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

[60] 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.

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

[62] 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.

[63] M.-P. Pelletier, M. Trépanier and C. Morency (2011). Smart card data use in public transit: a literature review. *Transportation Research Part C: Emerging Technologies* **19** (4), pp. 557–568.

[64] 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.

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

[66] 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*.

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

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

[69] H. Tenkanen, P. Saarsalmi, O. Järv, M. Salonen and T. Toivonen (2016). Health research needs more comprehensive accessibility measures: integrating time and transport modes from open data. *International Journal of Health Geographics* **15** (1), p. 23.

[70] 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.

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

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

[73] 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.

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

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

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

[77] 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.

[78] 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.

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

[80] 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.

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

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

[83] 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.

[84] 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.

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

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

[87] 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.

[88] 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.

[89] 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.

[90] 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.

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

[92] 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.

[93] 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*.

[94] 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.

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

[96] 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*.

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

[98] 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.

[99] 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.

[100] 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.

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

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

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

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

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

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

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

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

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

[110] 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.

[111] J.-P. Rodrigue, C. Comtois and B. Slack (2016). *The geography of transport systems.* Routledge.

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

[113] 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.