TU Delft

Delft University of Technology

Modeling Human Spatial Behavior Through Big Mobility Data

Wang, Y.

DOI
10.4233/uuid:510dd3e1-e5eb-4032-a785-c59df38f8c58

Publication date
2021

Document Version
Final published version

Citation (APA)
Wang, Y. (2021). *Modeling Human Spatial Behavior Through Big Mobility Data*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:510dd3e1-e5eb-4032-a785-c59df38f8c58

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Modeling Human Spatial Behavior through Big Mobility Data

Yihong Wang

This thesis was supported by the Netherlands Research School on Transport, Infrastructure and Logistics (TRAIL) and the Netherlands Organization for Scientific Research (NWO).
Grant code: 022.005.030.

Cover illustration: Huipeng Xu

Modeling Human Spatial Behavior through Big Mobility Data

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology,
by the authority of the Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on
Wednesday 23 June 2021 at 10:00 o'clock
by

Yihong WANG

Master of Science in Civil Engineering,
Delft University of Technology, the Netherlands
born in Shanghai, China

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	Chairperson
Prof.dr.ir. B. van Arem	Delft University of Technology, promotor
Prof.dr. H.J.P. Timmermans	Eindhoven University of Technology, promotor
Dr.ir. G. Homem de Almeida Correia	Delft University of Technology, copromotor

Independent members:

Prof.dr. Y. Susilo	Universität für Bodenkultur Wien
Prof.dr.ir. S. Rasouli	Eindhoven University of Technology
Prof.dr.ir. G.P. van Wee	Delft University of Technology
Prof.dr.ir. J.W.C. van Lint	Delft University of Technology

TRAIL Thesis Series no. T2021/19, the Netherlands Research School TRAIL

TRAIL
P.O. Box 5017
2600 GA Delft
The Netherlands
Phone: +31 (0) 15 278 6046
E-mail: info@rstrail.nl

ISBN: 978-90-5584-293-3 Copyright © 2021 by Yihong Wang

All rights reserved. No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without written permission from the author.

Printed in the Netherlands

Acknowledgement

Learning, as we know, includes supervised learning and unsupervised learning. This does not only apply to machine learning, but also to human learning.

It is not difficult to understand that PhD is a supervised learning process. Learning under the supervision of Gonçalo, Bart, Harry, and Erik was a great experience. I still remember the moment when I received the feedback from Gonçalo on a draft of my first work. Back then, I was frustrated with the large number of questions and tracked changes. After a few back-and-forths, I started building a predictive model in my head, which could somehow anticipate Gonçalo's reaction to every sentence and even every single word that I would write. Based on that, I was able to improve my writing. That was exactly when I realized the equivalence between supervised learning as a human and supervised learning as a machine.

This learning style is a microcosm that demonstrates how I learned during my PhD. Supervisors, in my humble opinion, are someones who help you label various kinds of data, or in other words, who help you define which is good and which is bad, from academic writing to searching research topics and solving research problems. I am truly grateful for learning so many things from all my supervisors.

The best thing I learned from Gonçalo is the spirit of ELI5 (Explain Like I'm Five) even in academia. In China, people have the stereotype that research has to be very complex. I was also mazed by this myth before coming to the Netherlands. Then I was so impressed by Gonçalo's lectures. After I started doing research with him, I have also learned that if something is still too complex to explain, it simply means I have not fully understood it yet. Bart is a leader and project manager in nature. His charisma has influenced me. One of his quotes from a progress meeting always comes to my mind when I need to make decisions: you do everything for a reason. Harry is such a pioneer and an OG in the field of mobility and machine learning. Machine learning is getting a lot of hype nowadays in the transportation field, and I feel people who always talk about machine learning, ironically, don't really understand it (like me). In contrast, people like Harry who have been working on this subject for decades would rather not tag their research with such trendy words. Because he understands the essentials, he is so visionary about the direction of cutting-edge research. Erik is the one who brought me to the world of big data and transportation. I still remember the excitement when I first saw the description of the Senegal project. It is a pity that

he has not been my promotor since the second year of my PhD, but I know I would not have had the chance to do all these without him.

Thanks also go to all my committee members: professors Rasouli, Susilo, van Wee and van Lint. Thanks for accepting our request and for spending the time reading my work, giving feedbacks and attending my defense. This is a great honor for me.

PhD is also a process of unsupervised learning. I sometimes felt just like an explorer while doing my own day-to-day research. I remember the days and nights spent exploring data in R driven by pure curiosity. Especially after I started working for a company, I am more grateful for those four years in which I was just trying to stretch the boundaries of a very small part of human knowledge without any utilitarian purpose. Shout out to TRAIL and NWO for providing the scholarships that give PhDs freedom to do the research they like.

Self-discovery is another thing that can only be learned in an unsupervised fashion. Honestly, I was like a Chinese hipster before coming to the Netherlands. Everyone who knew me in the same undergraduate program would possibly be shocked to see that I am now completing a PhD. On the other hand, in Europe, everyone just regards me as another random Asian dude. This makes me rethink about myself. Who am I? What is the uniqueness of myself? My favorite Chinese hip-hop artist *J-Fever* sang this: *Have you found the next pitch yet? Have you found a new crazy player to pass you the ball? Have you found a new reason to be crazy?* I am proud of myself being crazy about mobility research in the past years. Now I am looking forward to the new challenges. Stay tuned for the updates about my next pitch (https://github.com/bellowswang).

As an average reader who always only reads acknowledgement in a thesis, I noticed in this paragraph, PhDs would usually start dropping names, from which I can easily derive the local social network structure with real personal information. To follow the GDPR guidelines (I know this is a bad joke), I decide to aggregate and anonymize the information a bit. I want to thank all my colleagues in the Transport & Planning department. Back in China, I was a language-dependent social player specializing in wordplays and memes. In a different language, I was just like Rome without Caesar and a flight with no VISA, which made me realize the importance of transfer learning. Anyway, thanks for all the memories. I especially want to give special credits to my office roommates. Having coffee breaks together is always chill despite the taste of coffee. I want to highlight the trip to California in 2018 for IATBR and the trip to Washington, D.C. in 2019 for TRB. It is always a great experience to go to conferences all over the world with lovely colleagues. Let's keep connected (not only on LinkedIn)! I also want to thank my (relatively) new colleagues at Just Eat Takeaway.com, especially the SODA (Scoober Operations Data Analytics) team. Different from being a PhD working independently all the time (which I also enjoyed a lot), it has been a lot of fun (another kind of fun) working together with teammates.

I want to thank all my friends, my old friends and new friends, my friends in the Netherlands and outside the Netherlands. As my favorite American indie rock band *Kind of*

Like Spitting sang: *All my friends are brilliant.* Yes, it's you! I'm talking about you. When you read this, feel free to drop me a message and congratulate me! I want to especially thank my two bands as well: Feima (https://feima.bandcamp.com/album/half-city) and Animal Hierarchy (https://soundcloud.com/user-8714564). Let our music keep flowing.

Finally, I want to express my gratitude to my family. My parents are the best parents ever. Thank you for your constant support all the way! Let's make our life better and better. Thank you Huipeng. You are the best wife. We have been through so much together, and we made it! My thesis with the cover you designed is just perfect. Last but not least, I want to dedicate this thesis to my grandma. When you were alive, you might have been the oldest (and coolest) person in China who loved Harry Potter. This thesis on revealing human mobility is the Marauder's Map I made for you as a gift.

Yihong
Rotterdam, May 2021

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and scope

People move in time and space daily. Many questions can be posed about this phenomenon. Why do they travel? Why do they choose to visit a certain place? Why do they travel at a certain time? Why do they use the car over public transportation? Why do they follow a certain route? Travel behavior (interchangeably referred to as human mobility) research aims to answer all these questions, which can respectively be refined into the following dimensions of travel-related choices: activity type choice, location choice, time-of-day choice, transportation mode choice, and route choice (Rasouli & Timmermans, 2014; de Dios Ortúzar & Willumsen, 2011).

This thesis specifically focuses on location choice. Notwithstanding, activity type choice is inevitably in scope as well, because activity type choice is most likely a prerequisite for location choice. Different activity types result in different location choice sets (Arentze & Timmermans, 2004). For example, for daily work or home activities, most people do not have a choice because their home or work location is unique and fixed, and it has been determined on a longer-term basis, serving as anchor locations to perform other activities (Arentze et al., 2013). On the other hand, if people want to eat outside, there is a large choice set of places for them to visit (Yoon et al., 2012), and these types of activities are designated as flexible activities since they are flexible in time and space (Wang et al., 2016).

This thesis refers to the outcome of people's activity type and location choice in a mobility system or network, as human spatial behavior. Specifically, the outcome can be individual location choice for after-work flexible activities (Chapter 2), individual preferred destination type for flexible activities (Chapter 3), individual location choice for flexible activities (Chapter 4), or an aggregated origin-destination (OD) trip matrix (Chapter 5).

Human spatial behavior can be measured for each individual. We can observe the locations that a person visits for different activities in a certain time period (e.g., Yue

et al., 2014). Traditionally, researchers ask a group of respondents to report such information, as part of a so-called travel diary (e.g., Schlich & Axhausen, 2003). In recent years, such individual spatial traces can be tracked passively by new technology such as mobile phones and smart cards (e.g., Calabrese et al., 2015; Jiang et al., 2013; Pelletier et al., 2011). Since most people bring mobile phones and use smart cards in their daily lives, the mobility data collected in this way are likely to be big, thus designated as "big mobility data".

Compared to travel survey data, big mobility data can show a larger-scale picture of human spatial behavior in a city by tracking the spatial-temporal traces of many people (Demissie et al., 2015). For privacy concerns, big mobility data are sometimes prepared in an aggregated way, in terms of OD matrices (Caceres et al., 2013). This provides a macroscopic perspective to understand human spatial behavior (Sevtsuk & Ratti, 2010). From this perspective, activity type choices of individuals are aggregated into trip generation from each zone, and destination choices of individuals are aggregated into trip distribution between zones (Anas, 1983). This thesis aims to use big mobility data, either aggregated or disaggregated, to contribute to the understanding of human spatial behavior.

An individual's location choice for performing an activity can largely be explained by three types of factors: individual-specific factors, location-specific factors, and accessibility factors (Horni, 2013). Individual-specific factors are related to travelers' attributes such as socioeconomic status. Location-specific factors describe the characteristics of a location and/or its surrounding urban environment. Accessibility factors generally measure the extent to which transportation and land-use systems enable individuals to reach destinations (Geurs & Van Wee, 2004). All these factors are usually assumed to be observable and serve as explanatory variables in location choice models.

Using disaggregated mobility data, researchers attempt to understand the importance of a certain location and/or accessibility factor given personal characteristics from a choice theory perspective (Koppelman, 2007). The importance is regarded as an unknown parameter in a discrete choice model and estimated by fitting historical individual travel data. Using aggregated mobility data, individual-specific factors are averaged out, and a gravity model is commonly used to explain trip distribution based on location-specific factors and accessibility factors (Hansen, 1959). In summary, discrete choice models and gravity models are the two most common types of models to understand human spatial behavior. Apart from these types of models, this thesis explores and expands the body of knowledge on new spatial behavior models that are more appropriate for big mobility data.

Transportation services, planning and policies can shape human spatial behavior (Fox, 1995). For example, if a metro is operated overnight, night event locations might attract more people. To plan and operate better transportation services, decision-makers first want to know the current picture of human mobility in their systems or networks so that they can conduct ex-post evaluations. When only survey data are available, they first need to estimate a model of spatial behavior using a collected sample, and then

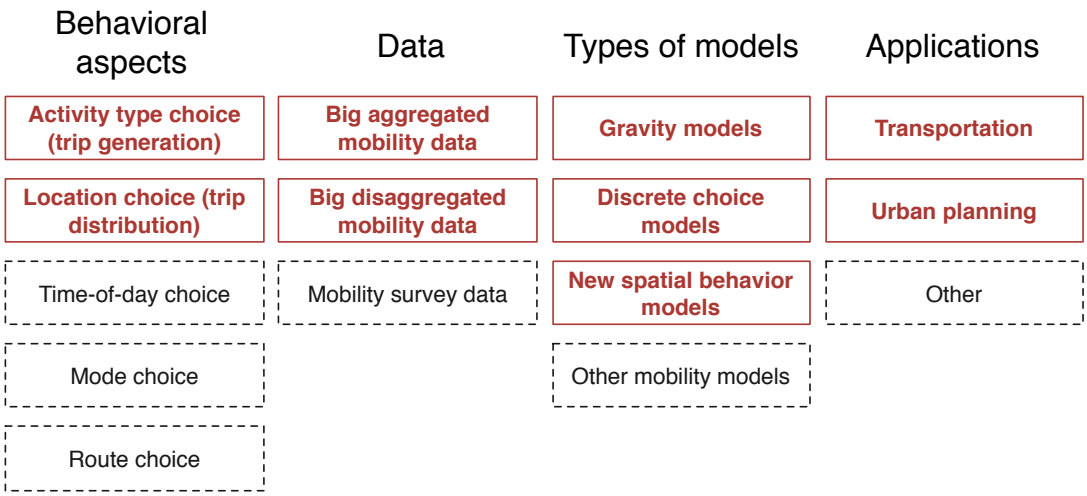


Figure 1.1: The overview of human mobility research and the scope of this thesis highlighted in red.

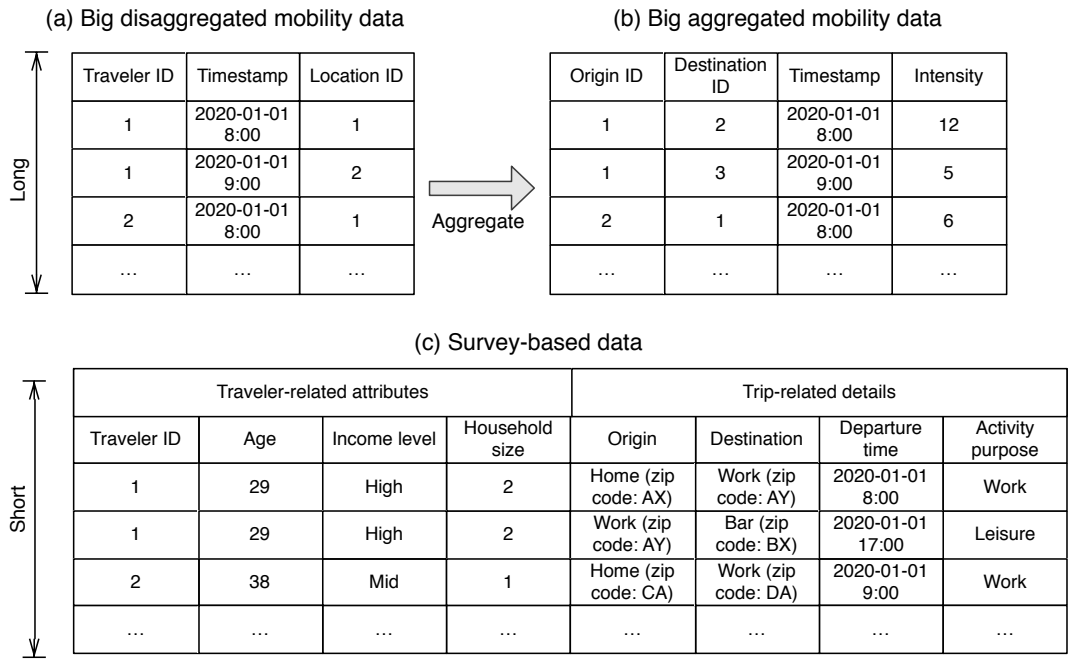
extrapolate to the whole population (Tolouei et al., 2017). Nowadays, some big mobility data themselves can already tell an almost complete story about historical mobility patterns (Sevtsuk & Ratti, 2010). Moreover, decision-makers are also eager to foresee how their decisions would finally impact human mobility. For this purpose, they need to apply spatial behavior models in what-if scenarios, and predict behavioral responses to policy scenarios.

Spatial behavior models can also be applied for purposes other than transportation and urban planning, including but not limited to controlling spread of diseases (Balcan et al., 2010) and socioeconomic well-being (Pappalardo et al., 2015), which are out the scope of this paper. As a summary, Figure 1.1 presents the scope of this thesis, highlighted in red, as well as its position in the larger realm of mobility research.

1.2 Motivation

1.2.1 Data

Big mobility data vs. survey-based data (called as "small data" in Chen et al., 2016) has been a topic of long-time debate in human mobility research. Big data are intuitively better than relatively "small" survey data but this is not always the case (Bonnel et al., 2015). This thesis argues that in most cases, big mobility data are only big in terms of the number of samples, but not big in terms of the number of features; survey-based data are exactly the opposite. Big mobility data contain a large number of travelers and trips but little is known about each traveler and trip, not to mention that sometimes they have to be aggregated. On the other hand, survey-based data, despite reporting only a small group of respondents, tend to include abundant features about each traveler,



(a) Big disaggregated mobility data

Traveler ID	Timestamp	Location ID
1	2020-01-01 8:00	1
1	2020-01-01 9:00	2
2	2020-01-01 8:00	1
…	…	…

(b) Big aggregated mobility data

Origin ID	Destination ID	Timestamp	Intensity
1	2	2020-01-01 8:00	12
1	3	2020-01-01 9:00	5
2	1	2020-01-01 8:00	6
…	…	…	…

(c) Survey-based data

Traveler-related attributes				Trip-related details			
Traveler ID	Age	Income level	Household size	Origin	Destination	Departure time	Activity purpose
1	29	High	2	Home (zip code: AX)	Work (zip code: AY)	2020-01-01 8:00	Work
1	29	High	2	Work (zip code: AY)	Bar (zip code: BX)	2020-01-01 17:00	Leisure
2	38	Mid	1	Home (zip code: CA)	Work (zip code: DA)	2020-01-01 9:00	Work
…	…	…	…	…	…	…	…

Figure 1.2: A pseudo-example of typical big mobility data vs. survey-based data.

such as age, and about each trip, such as trip purpose (Kwan, 2016). Assuming that each row represents one sample and each column represents one feature, big mobility data should have better been described as long and thin, and "small" survey-based data should have better been described as short and wide.

The difference between typical big mobility data vs. survey-based data is illustrated in Figure 1.2. Most big disaggregated mobility data record the spatial-temporal traces (i.e., locations and timestamps) of many individuals in a certain geographic area in a time frame (Çolak et al., 2015), whilst big aggregated mobility data present the intensity of spatial interactions between every two locations per time slot (Deeva et al., 2019). Survey-based data are often disaggregated, and focus on one geographic area and a certain period. Different from big mobility data, these data not only include the accurate origin, destination, departure time and arrival time of each trip made by the respondents, but also further details about each trip as well as the attributes of each respondent (Collia et al., 2003).

Big disaggregated mobility data cannot be long and wide at the same time mainly because of privacy concerns. For example, mobile phone traces cannot include the personal information of a certain mobile phone user (De Montjoye et al., 2013). Survey-based data cannot become longer because they are expensive to collect. Also, due to the cost of data collection, survey-based data are usually not updated (Alexander et al., 2015).

The obsolescence of survey-based data was not a serious problem in the days when they were mainly used for long-term transportation planning purposes. Today, in this hyper-connected, technological world, mobility data are being consumed by more par-

ties, including not only planning authorities but also more retail and mobility companies (e.g., Cohen et al., 2016; Timmermans, 1993). All need the most affordable and updated mobility data to make more timely decisions in a cost-efficient way. It has thus become relevant to explore the use of big mobility data, especially in terms of how to leverage their strength (i.e., being long) and avoid their shortcoming (i.e., being thin or being aggregated).

1.2.2 Models and applications

After collecting survey-based mobility data, the next step is to estimate the population's behavior based on the samples. For example, trip frequency per age group can be inferred. Since survey-based data include the attributes of each traveler, it is feasible to extrapolate from the sample to the population, as long as the distribution of each attribute is known at the population level. In many cases, mobility surveys can result in a general report that summarizes travel behavior of the population (Collia et al., 2003; Lu & Gu, 2011). Also, researchers can estimate travel behavior models using small-size survey data and then apply the models to a synthetic population, so that they can estimate a full picture of current mobility patterns (Ziemke et al., 2019).

Comparably, more efforts have to be made to extract spatial behavior information from big mobility data. As illustrated in Figure 1.2, big mobility data mostly reveal only two elements of human spatial behavior: location and time. However, neither of them is necessarily accurate in big mobility data, which are collected passively and thus not meant for mobility-related purposes in its nature. For example, many spatial-temporal traces of a traveler can be left in mobile phone data, but mobility researchers want to distinguish the real activity locations from the other pass-by places (e.g., Zheng et al., 2009). Certain techniques are therefore necessary to extract real trip information, including origin, destination (e.g., Alexander et al., 2015) and departure time (e.g., Bwambale et al., 2019).

One might think that big data are more representative of the population. This could be true in some cases but sometimes it is even more difficult for big mobility data to represent accurately the population. For example, mobile phone data could be biased if they are only from one telecommunication provider (Zhao et al., 2016b). Social media check-in data, as a trendy mobility source, have been criticized for being biased towards young people (Huang & Wong, 2016). Consequently, the estimation of travel demand would be negatively impacted. In recent years, extensive research has attempted to overcome the aforementioned issues in order to allow the possibility of using big mobility data to provide an accurate overview of spatial behavior and travel demand (Munizaga & Palma, 2012; Alsger et al., 2015; Iqbal et al., 2014; Demissie et al., 2016). Although estimation is not the main focus of this thesis, it is a task that cannot be bypassed before understanding spatial behavior and travel demand (as illustrated in Figure 1.3). This thesis reviews the existing methods and adapts them to fit in our specific cases.

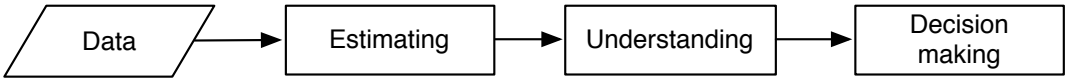


Figure 1.3: The process of transforming mobility data into informed decisions.

To take a further step, it is also worth building predictive models based on the estimated mobility information because they can help understand spatial behavior and make informed decisions (Ben-Akiva et al., 1996). For example, one can estimate the current OD trip matrix, but without understanding it, this cannot directly result in any decision, unless it is possible to predict a new OD matrix given a different transportation network layout. Survey-based data are convenient for building a predictive spatial behavior model because of the large number of features. For example, location choice patterns per age group can be learned, and predictions can be made accordingly (Arentze et al., 2013). However, big mobility data do not fit in this approach since few features are available. While differences in spatial behavior can still be observed, they are difficult to explain because it is difficult to know who the travelers are and why they travel (Calabrese et al., 2013).

In summary, big mobility data are favorable for being cost-efficiently, up to date and promising especially in terms of sample size, but given their very nature (i.e., being thin or being aggregated), it is still cumbersome to use them for understanding human spatial behavior. This thesis aims to contribute to filling this gap by exploring the answers to the main research question which is formulated as follows:

To what extent, and how, can big mobility data foster the understanding of human spatial behavior?

1.3 Contributions

1.3.1 Scientific contributions

Most existing spatial behavior models are theory-based. Typical examples include discrete choice models based on the utility maximization theory and gravity models based on the physics theory. Those models inherently require input data to be sufficiently wide to include features supporting their respective theories. For example, to account for individual discrete choice, there should be data related to each component of utility and individual characteristics. Fitting such theory-based models with long-and-thin big data is possible, but it would be a lose-lose situation: theory-based models would be weakened by the lack of features, and patterns latent in large samples would not be fully explored because of the constraints of theories. One potential solution is to feed theory-based models with an expanded dataset. The other potential solution is

to use data-driven models, which essentially make less strong assumptions about the nature of the data distributions than theory-based models (Murphy, 2012).

The scientific contribution of this thesis consists of two main strategies adopted to answer the research question. The first principal strategy is to make long and thin data wider. This strategy has led to the following publications, which correspond to Chapter 2 and 3 respectively:

Wang, Y., Correia, G.H.A., de Romph, E., & Timmermans, H.J.P. (2017). Using metro smart card data to model location choice of after-work activities: An application to Shanghai. Journal of Transport Geography, 63, 40-47.

Wang, Y., Correia, G.H.A., van Arem, B., & Timmermans, H.J.P. (2018). Understanding travellers' preferences for different types of trip destination based on mobile internet usage data. Transportation Research Part C: Emerging Technologies, 90, 247-259.

Since lack of features is the biggest obstacle for big mobility data to explain human spatial behavior, attempts are made to generate proxy variables for traveler segmentation and trip characterization, from either big mobility data themselves (Chapter 2) or external datasets (Chapter 3). The addition of proxy variables for each traveler and each trip can enhance the understanding of human spatial behavior. This principal strategy results in the following methodological contributions:

- Adapting the existing algorithms to our case study to detect home and work stations of metro travelers from disaggregated smart card data (Chapter 2).
- Proposing to use home and work stations as proxy variables to distinguish behavior heterogeneity 2).
- Building a discrete choice model with the addition of the proposed proxy variables to model after-work activity location choice in a metro network by using disaggregated smart card data (Chapter 2).
- Building a clustering algorithm to distinguish the functions of urban areas based on point of interest (POI) data and using the results to label trip destinations extracted from disaggregated mobile phone traces (Chapter 3).
- Testing the hypothesis that one's preferred destination types are related to one's preferred mobile internet content, extracted from mobile internet usage data (Chapter 3).

The second principal strategy takes a new and groundbreaking approach, inspired by the collaborative filtering algorithms that are commonly used to model user preferences in recommendation systems (Koren et al., 2009). This strategy has led to the following under-review articles, which correspond to Chapter 4 and 5 respectively:

Wang, Y., Correia, G.H.A., van Arem, B., & Timmermans, H.J.P. (2020). Exploring a neighborhood-based collaborative filtering approach to modeling location preferences for flexible activities through metro smart card data. Journal of Transport Geography, submitted.

Wang, Y., Correia, G.H.A., van Arem, B., & Timmermans, H.J.P. (2020). A matrix factorization approach to modeling trip generators and their interactions. Travel Behaviour and Society, submitted.

Without using any specific proxy variables, Chapter 4 and 5 implement data-driven methods, which only rely on empirical observations about many people, and do not require imposing any theory-based prior assumptions about the mechanisms of human spatial behavior. The intuitive reason why this approach might work is that historical spatial behavior itself can indicate some heterogeneity between individuals within a given group of travelers and thus help make predictions about their future behavior. This principal strategy results in the following methodological contributions:

- Building a neighborhood-based collaborative filtering algorithm to model location preferences for non-work activities in a metro network by using disaggregated smart card data (Chapter 4).
- Building a Poisson factorization algorithm to model spatial interactions in a metro network by using aggregated smart card data (Chapter 5).

1.3.2 Practical contributions

As pointed out in Figure 1.3, the process of transforming mobility data into informed decisions includes three stages: estimating, understanding, and decision making. All the analyses that were conducted in this thesis cover the first two parts, and especially contribute to the second part. Urban authorities, mobility companies and retail companies can follow our approaches to estimating and understanding human spatial behavior using their own big mobility data. For example, a public transportation operating company stores massive mobility data of its services, and it can freely apply our methods to extract mobility information and understand the spatial behavior of its users.

Although this thesis does not include the part of decision making, it is promising to do so based on the understanding of human spatial behavior. For example, Chapter 2 builds a location choice model for after-work activities in a metro network. Urban planners can further use this model as a starting point to optimize the development of shopping areas around metro stations.

A side note on the practical contributions is about data privacy issues. Since the mobility data might reveal highly sensitive personal information, the use of big mobility data, especially in disaggregated form, could be restricted for research and analysis

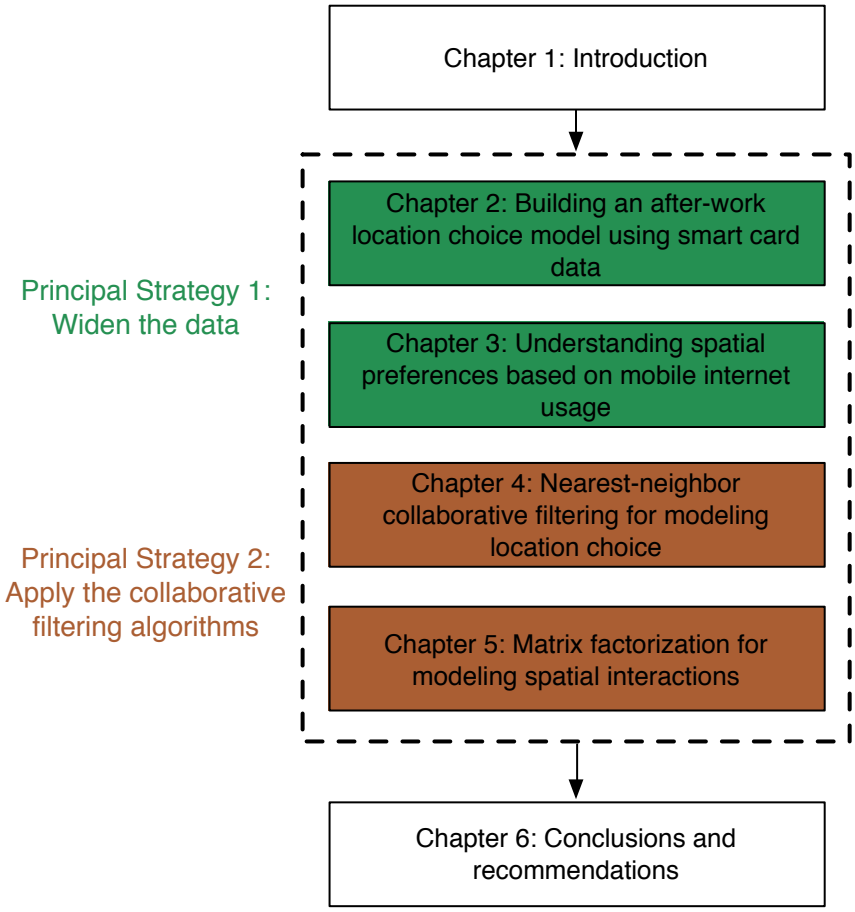


Figure 1.4: The outline of this thesis.

(Giannotti & Pedreschi, 2008). Given local regulations, the available granularity of big mobility data varies from case to case. Considering this issue, this thesis proposes various methods, which can deal with not only big disaggregated mobility data, but also big aggregated mobility data.

1.4 Thesis outline

The remainder of the thesis is organized as follows.

Chapter 2 takes the advantage of the long nature of the metro smart card data of Shanghai, China to detect the stations that are closest to home and work locations of each metro traveler. In most cases, if a traveler always leaves one station at the beginning of a day and returns to this station at the end of the day, this station is likely to be associated with this traveler's home location, thus named as home station. Such socio-geographic information can help characterize trip purposes, and as a result, those trips for after-work activities are especially distinguished in the case study. Although personal attributes are not explicitly provided in big mobility data, detected home and

work stations might be able to indicate some behavior heterogeneity among travelers. For example, metro travelers living in a more residential area might share some common characteristics. Based on this idea, Chapter 2 adds two proxy variables to distinguish travelers living/working in the different types of areas, and the variables are found to help enhance the prediction accuracy of a discrete choice model accounting for after-work activity location choice.

In the era of mobile internet, users generate not only spatial-temporal traces but also internet browsing traces. Chapter 3 fuses mobile phone traces with a special external dataset: the mobile internet usage data of the same users. The objective is to understand users' spatial preferences based on their mobile internet usage, which also serves as a proxy variable for personal attributes. Moreover, POI data, which record the coordinates of each POI, are also used as an external source to characterize trip destinations, based on a clustering algorithm.

Chapter 4 argues that the previous approaches rely on theory-based assumptions and thus proposes a data-driven approach under a more flexible assumption: past behavior itself can reflect the heterogeneity in the population and be further used as a reference to predict future behavior. Specifically, this chapter introduces an algorithm called neighborhood-based collaborative filtering, which finds the so-called neighbors of a traveler. Instead of being geographically close to each other, the neighbors in this definition are similar in terms of past spatial behavior.

Chapter 5 continues the data-driven strategy by implementing the other main-stream collaborative filtering algorithm: matrix factorization. Instead of factorizing a traveler-location frequency matrix using big disaggregated mobility data, this chapter considers data privacy issues and proposes a Poisson factorization method, a variant of the classical matrix factorization algorithm, to model aggregated spatial behavior, in terms of a location-location frequency matrix (i.e., spatial interaction matrix or OD matrix).

Finally, Chapter 6 presents the conclusions of the thesis and recommendations for future research.

The outline of the thesis is shown in 1.4.

Chapter 2

Building an after-work location choice model using smart card data

---

Chapter 1 identified a problem in human spatial behavior modeling using big mobility data: the absence of features accounting for behavioral heterogeneity. A straightforward solution is to use proxy variables for personal attributes. This chapter specifically investigates the possibility of using socio-geographic status as a proxy for personal attributes to model after-work location choice; i.e., given a metro commuter's home and work locations, the question is: can we predict where this person would visit after work? To solve this problem, a discrete choice model is estimated using metro smart card data from Shanghai, China. The model could further serve as a tool to help retail companies locate their businesses optimally and help urban decision makers plan transport networks and land use more reasonably.

The chapter is based on the following publication:

Wang, Y., Correia, G.H.A., de Romph, E., & Timmermans, H.J.P. (2017). Using metro smart card data to model location choice of after-work activities: An application to Shanghai. Journal of Transport Geography, 63, 40-47.

Abstract

A location choice model explains how travelers choose their trip destinations especially for those activities which are flexible in space and time. The model is usually estimated using travel survey data; however, little is known about how to use smart card data (SCD) for this purpose in a public transport network. Our study extracted trip information from SCD to model location choice of after-work activities. We newly defined the metrics of travel impedance in this case. Moreover, since socio-demographic information is missing in such anonymous data, we used observable proxy indicators, including commuting distance and the characteristics of ones home and workplace stations, to capture some interpersonal heterogeneity. Such heterogeneity is expected to distinguish the population and better explain the difference of their location choice behaviour. The approach was applied to metro travellers in the city of Shanghai, China. As a result, the model performs well in explaining the choices. Our new metrics of travel impedance to access an after-work activity result in a better model fit than the existing metrics and add additional interpretability to the results. Moreover, the proxy variables distinguishing the population seem to influence the choice behaviour and thus improve the model performance.

Keywords: Public transport; smart card data; location choice modelling; discrete choice model; demand forecast; transport planning.

2.1 Introduction

Travel behaviour is becoming more diverse and complex especially in large metropolitan areas. One of the most significant changes is that non-commuting travel demand takes a larger share than ever before (Lu & Gu, 2011). Therefore, the task of observing and analysing non-commuting travel demand is becoming important today. This task is not only relevant for transport planners to better understand movements of travellers, but also for service and retail business planners to understand where people would like to consume and where their customers come from (Sivakumar & Bhat, 2007). Moreover, economists regard the accessibility to non-commuting activities as an important indicator to reflect quality of life (Nakamura et al., 2016; Suriñach et al., 2000). These relevant perspectives have led the transportation research field to expand its scope to topics like accessibility (Dong et al., 2006), social exclusion (Schönfelder & Axhausen, 2003), subjective well-being (De Vos et al., 2013), etc., in addition to traditional transport problems particularly focusing on network levels of service.

To cope with the increasing non-commuting demand, the usage of public transport (PT) to access retail and service facilities has been encouraged in many cities due to the concentration of people (Castillo-Manzano & López-Valpuesta, 2009; Ibrahim & McGoldrick, 2017). Urban decision makers need to know where large recreational centres

should be located and how PT network should be planned to meet the considered objectives. Answering these questions requires the prediction of non-commuting OD matrices in many what-if scenarios, based on the understanding of peoples activity-travel behaviour including, but not limited to, location choice. A relevant and interesting perspective is the activity-based travel demand modelling, which focuses on individuals and regards travelling as the result of the need to participate in activities (Rasouli & Timmermans, 2014). However, few studies have adopted this methodology focused on PT network. In this paper, we aim to fill this gap by using new available travel demand data sources, namely, smart card data (SCD). We focus on travel demand of after-work activities since it is a significant part of non-commuting travel demand especially on weekdays (Demerouti et al., 2009). Our research can also be regarded as a complement to the existing research that uses SCD to study commuting patterns (Ma et al., 2017; Zhou et al., 2014).

Compared to traditional mobility survey data, SCD have several advantages and disadvantages to reveal how people travel by PT (Bagchi & White, 2005; Pelletier et al., 2011). Firstly, collecting such data is more efficient, saving both time and money, compared to large-scale surveys. Secondly, SCD usually correspond to a larger sample and the observations can be longitudinal in time (Morency et al., 2007). On the other hand, trip purpose is difficult to obtain in SCD and needs to be estimated using other methods (Devillaine et al., 2012; Kuhlman, 2015; Long et al., 2012). In some cases, destination information needs to be estimated as well because some PT networks do not request a check-out (Trépanier et al., 2007). The very relevant personal socio-demographic information is most of the times not available for confidentiality reasons which decreases the possibility to do a more thorough analysis of particular behavioural traits of the population (Pelletier et al., 2011).

The advantages of using SCD have allowed researchers to obtain more accurate estimates of transit demand, which have led to many applications. Using the data collected during 277 consecutive days, Morency et al. (2007) examined the variability of transit use. Some studies proposed to cluster and classify the regularity of transit travel patterns by mining SCD (Goulet-Langlois et al., 2016; Ma et al., 2013). Estimating origin-destination (OD) transit trip matrices is a usual application of SCD (Munizaga & Palma, 2012). It can further serve as a fixed input to passenger flow assignment (Sun et al., 2015), OD flow visualization (Liu et al., 2009; Long et al., 2012) and any other post hoc analysis, such as commuting efficiency assessment (Zhou et al., 2014). However, only a few attempts have been made to use SCD to build explanatory trip distribution or location choice models, in order to predict the OD matrices as a result of the changes made to transport systems and land use. One example is the gravity model developed by Goh et al. (2012) to understand aggregate commuting OD flows by metro. We believe that not only the characteristics of SCD but also the research objective in our study is a better fit for a disaggregate activity-based travel demand modelling framework.

In this study, we use SCD to model location choice of after-work activities. The in-

novation of our approach firstly lies in the creation of new metrics to model travel impedance in location choice of after-work activities. Secondly, this is the first time that proxy variables, which can be observed in anonymous SCD, are used to capture some interpersonal heterogeneity in order to explain the difference of their location choice behaviour. Thanks to the Shanghai Open Data Apps (SODA) contest[1], a full-population dataset of one-month PT smart card transaction records for the city of Shanghai (China) was made available, allowing us to explore this methodology in a large-size real-world case scenario.

This paper is organized as follows. First, the methodology is described. Then, the data of Shanghai is further explained. Following that, we present the application of our method. In the final section, we take conclusions and point out directions for future research.

2.2 Methodology

We start by defining the scope to which our methodology can be applied. The method can be applied in a metro network composed of stations with services connecting them, where the automated fare collection system forces travellers to check in and check out at the stations where they board and alight respectively. Therefore, the following information of each trip is available through SCD: anonymous identity (ID) of the user, IDs of boarding and alighting stations and timestamp. A trip is defined to start from an origin station near which the previous activity has been finished, and end at a destination station where the next activity will take place. In our case, the recorded boarding and alighting stations are not necessarily an origin or a destination station of a trip. In other words, a trip including any transfers should not be regarded as two separate ones. Moreover, a daily trip chain is the ordered set of trips done by an individual within one day.

2.2.1 Detecting commuters

Several studies have been performed on the detection of commuters as well as their home and workplace stations from SCD (Chakirov & Erath, 2012; Long & Thill, 2015). By recurring to travel survey data, researchers have either predefined the rules or trained the models to predict if a smart card user is a commuter and if the purpose of a PT trip recorded in SCD is home, work or other, based on several observed factors, such as activity start time. In our method, we used a similar principle for activity identification, but due to the unavailability of travel survey data, we predefined the rules with the parameters identified in the literature. We used the following rule applied by Long et al. (2012) to determine ones home station: any boarding station of the first trip

[1] http://soda.datashanghai.gov.cn/ (retrieved date: November 21st, 2015)

done by an individual on a weekday was defined as a so-called candidate home station of this individual, and the station appearing most frequently as a candidate home station during the observed period was defined as the definitive home station of this individual. There could be more than one station appearing most frequently. In such cases, Long et al. (2012) compared the land use around the stations and assigned the station in a more residential environment to be the definitive home station.

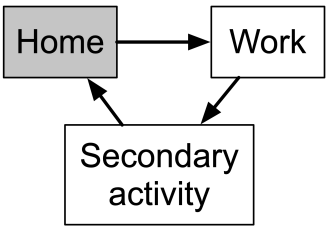
In SCD, activity duration can approximately be regarded as the time gap between a check-out and the subsequent check-in at the same station when the access and egress mode is walking. If the activity duration of visiting a station was longer than 6 hours on a weekday, we identified the station as a so-called candidate workplace station. Long et al. (2012) selected this parameter based on the travel survey data from Beijing, China, and thus we think that it is the best reference for our study of Shanghai despite the differences between the two cities. Next, the station appearing most frequently as a candidate workplace station during the observed period was defined as the definitive workplace station. If there were more than one station appearing most frequently, we calculated for each station the distance from home multiplied by the frequency of visits during the observed period, as suggested by Alexander et al. (2015), and the station with the largest product was defined as the definitive workplace station.

Commuters were defined as those who had both detected definitive home and workplace stations. Due to access and egress, home and workplace stations are not, in many cases, the real locations of home and workplace but can be regarded as proxies for those, especially when the access and egress mode is walking. One drawback of our method is that those commuters who have multiple home or workplace stations or have flexible working hours are difficult to detect. If necessary and possible, we recommend a more flexible approach relying on travel survey data. However, this step is not the main focus of our work, and our current method using the parameters identified in the literature is sufficient to detect a great number of commuters whom we can study regarding their after-work station choice behaviour.

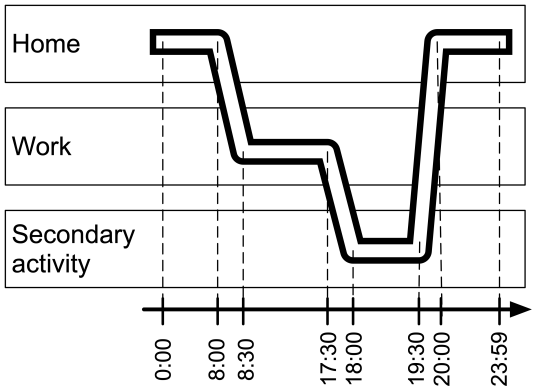
2.2.2 Extracting individual daily metro trip chains

We assume that within one day, travellers do an activity between every two consecutive trips, and the purpose of this activity can be estimated based on the check-out station of the former trip and the check-in station of the latter. If they are the same one, the purpose can be classified into home, work or secondary activity dependent on whether the station is the home station, the workplace station or neither for that individual; if they are different due to the interim unobservable movement by using other modes, we do not classify any activity purpose. Note that the first activity on one day is dependent only on the check-in station of the first trip, and the last activity is dependent only on the check-out station of the last trip.

The diagram of an individual daily metro trip chain starts in the first activity within a day, represented as a node, connected by an edge representing the trip to the second



(a) The diagram of a daily metro trip chain.



(b) The program of this trip chain.

Figure 2.1: An example of an individual daily metro trip chain.

activity, connected sequentially until the last activity. An example is shown in Figure 2.1, where each activity is labelled with its type and the grey box indicates where the chain starts. The commuter first travels from the home station to the workplace station at 8:00 and stays at the workplace station until 17:30. After staying at another station for 90 minutes, this person checks out there and travels back home.

2.2.3 Modeling station choices for after-work activities

In this paper, we focus on modelling station choice of metro commuters for after-work activities. Location choice involves a trade-off between attractiveness and travel impedance. We assumed that the attractiveness of a station for after-work activities is time-invariant. Travel impedance is a function of PT travel time, PT network distance, PT costs and number of PT transfers. In existing location choice models, there were three ways to model travel impedance to perform a secondary activity in a trip chain. The traditional way was to consider only the impedance of travelling between the activity location and home (Arentze & Timmermans, 2004). However, Arentze & Timmermans (2007) found that this measurement would result in the overestimation of the impedance between locations of activities within trip chains, and they proposed the concept of detour travel impedance:

$$DT_s = d(O_s, s) + d(s, D_s) - d(O_s - D_s) \tag{2.1}$$

In this equation, O_s is the origin of the trip to a candidate location s for the secondary activity, and D_s is the destination of the trip from s. $d(x,y)$ is the travel impedance from x to y.

Despite the wide use of this concept in existing travel demand models, such as MATSim (Horni, 2013), a disadvantage of this method is that it is not very sensitive in differentiating between distance from workplace or to home. Thus, while the previous definitions were adequate in the specific contexts of those studies, for our problem, it may be better to account for the effect of proximity to workplace vs. home. We defined the new metrics by complementing the detour impedance DT_s with a new variable R_s:

$$R_s = d(s,D_S) - d(O_s,s) \tag{2.2}$$

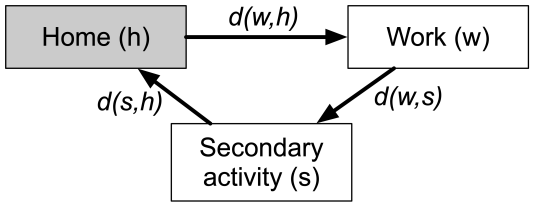
Table 2.1 summarizes the three possible ways to model travel impedance to perform an after-work activity in a trip chain. h, w and s represent home station, workplace station and candidate station for an after-work activity respectively, and the former two are respectively equivalent to the succeeding activity location D_s and the preceding activity location O_s in our specific case.

Although we focus on a metro network, attention should be paid to other modes like the access and egress to trips made in the metro network. In this study, we only model the trips to perform after-work activities with walking as access and egress, and we assume that the generalized travel cost of walking access and egress is minor compared to the main part of the metro trip.

The characteristics of activities (i.e., activity start time and activity duration) can be inserted in the model to describe contexts of choice occasions. The underlying assumption, in line with existing travel demand models (Balmer et al., 2008), is that people have already generated their activity schedules before making location choices. Attributes related to individuals are generally missing in SCD; however, in our study, we proposed to use commuting distance and characteristics of home and workplace stations as proxies for the attributes of the travellers. Aggregating the number of people living and working near each station can help identify whether a station is categorized into a mainly residential area or a mainly commercial area (Liu et al., 2009). This can serve as a way to characterize each travellers home and workplace stations.

Considering that choice making may also rely on the previously made choices, we include the effect of last choice feedback (i.e., first-order state dependence) in our model. Following the approach of Danalet et al. (2016), we estimate the model where the previous choice can be assumed to be strictly exogenous to the estimation. Danalet et al. (2016) also addressed a more advanced approach to deal with the initial conditions problem and related endogeneity bias in estimation. However, the consideration of these issues is beyond the scope of our paper. For the same reason, we do not consider time-variant attributes of alternatives and unobserved inter-individual and intra-individual response heterogeneity.

Table 2.1: Three ways to consider travel impedance in the choice of a location for an after-work activity.

	Existing metrics		**New metrics**
Measurement	Home-based impedance $d(s,h)$	Detour impedance $DT_s = d(w,s) + d(s,h) - d(w,h)$	Detour impedance DT_s and proximity to workplace vs. home $R_s = d(s,h) - d(w,s)$
Reference	Arentze & Timmermans (2004)	Arentze & Timmermans (2007); Horni (2013)	The approach in this study
Diagram			

We used a discrete choice model to explain the station choice for after-work activities with the referred impedance structures in our study. Consider that an individual user u in the network of the study area is associated with the home station h_u and the work-place station w_s, where $h_u, w_u \in \boldsymbol{N}$, and $\boldsymbol{N}$ is the set of metro stations in an area. In addition, u is observed to have a set of choice occasions $\boldsymbol{J}_u$ over time. The choice set of the destinations for after-work activities is denoted as $\boldsymbol{S}_{uj} = \boldsymbol{R}_{uj} \setminus \{h_u, w_u\}$, where $\boldsymbol{R}_{uj}$ is the reachable subset of N for u on choice occasion j. $\boldsymbol{R}_{uj}$ was calculated based on the following space-time constraints: (1) a commuter should not leave work earlier than the work schedule allows; (2) a commuter should not miss the last metro back home; (3) given the previous constraints, travel times to reach an after-work activity should not affect the activity start time and the activity duration. For each individual, we calculated the earliest time of departure from work during the observed period as the threshold to apply the first constraint. The timetables of the metro line were used to apply the second constraint. Travel time between every two stations can be calculated by averaging over the trips according to the SCD.

The deterministic part of the utility function for an alternative $s \in \boldsymbol{S}_{uj}$ on choice occasion $j \in \boldsymbol{J}_u$ of decision maker u in one month is the following:

$$\begin{aligned} V_{usj} = Z_s[\alpha + \sum_m (\delta_m X_{um}) + \sum_n (\phi_n C_{ujn})] \\ + \sum_k \{T_{usk}[\beta_k + \sum_m (\omega_{km} X_{um}) + \sum_n (\eta_{kn} C_{ujn})]\} \\ + \gamma SAME_{usj} \end{aligned} \tag{2.3}$$

Z is station attractiveness measured in terms of number of points of interest (POI). T is travel impedance. X is proxy variable for user-specific attributes. C is activity context. $SAME$ is about previous choice. $\alpha + \sum_m (\delta_m X_{um}) + \sum_n (\phi_n C_{ujn})$ is a function representing the preference for station attractiveness Z_s, and $\beta_k + \sum_m (\omega_{km} X_{um}) + \sum_n (\eta_{kn} C_{ujn})$ is a function representing the preference for reducing travel impedance T_{usk}. Both functions incorporate the effects of user-specific attributes X_{um} and activity characteristics C_{ujn} on taste variation. Therefore, the preferences vary across individuals and choice occasions (Sivakumar & Bhat, 2007). The descriptions of all variables and parameters are presented in Table 2.2, and the specific indicators of T_{usk}, X_{um} and C_{ujn} are summarized in Table 2.3. The possible values of $SAME_{usj}$ under different conditions are given in the following equation:

$$SAME_{usj} = \begin{cases} 1 & \text{if individual } u \text{ chose station } s \text{ on choice occasion } j-1 \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

Regarding the random part of the utility function, we used the spatially correlated logit model proposed by Bhat & Guo (2004) to consider the effect of spatial correlation

Table 2.2: Variables and parameters in the deterministic utility function.

Parameters		**Variables**	
γ	Preference for maintaining the previous choice	$SAME_{usj}$	Variable indicating the previous choice feedback
α	Baseline preference for attractiveness of station s	Z_s	Attractiveness of station s
β_k	Baseline preference for reducing the type k travel impedance	T_{usk}	The type k travel impedance associated with home and workplace station of individual u and station s
δ_m	The extent of the preference for attractiveness of station s that can be captured by the attribute m of travelers	X_{um}	Variable for the attribute m of individual u
ϕ_n	The extent of the preference for attractiveness of station s that can be captured by the characteristic n of activities	C_{ujn}	Variable for the characteristic n of the activity performed by individual u on choice occasion j
ω_{km}	The extent of the preference for reducing the type k travel impedance that can be captured by the attribute m of travelers		
η_{kn}	The extent of the preference for reducing the type k travel impedance that can be captured by the characteristic n of activities		

between adjacent stations on the metro network. This is a cross-nested logit model (Train, 2009) with two characteristics: (1) it is a paired combinatorial logit model (Koppelman & Wen, 2000), and each paired nest includes a station and one of its adjacent station; (2) it defines the allocation parameters that reflect the degree to which each alternative belongs to each nest. The probability of choosing an alternative can be calculated in a closed-form expression, where the dissimilarity parameter ρ ($0 < \rho \leq 1$) is designed to be equal across all paired nests and capture the general correlation between adjacent stations. There is no correlation between adjacent pairs of stations when $\rho = 1$, and the correlation increases as ρ decreases. In addition to the parameters in the deterministic part of the utility function, we need to estimate ρ as well. More details about the spatially correlated logit model can be found in the work by Bhat & Guo (2004).

Table 2.3: Indicators of travel impedance, user-specific attributes and activity characteristics in the utility function.

Variables		**Specific indicators**
Travel impedance variables	Home-based impedance	$T_{us1} = d(s,h)$
	Detour impedance	$T_{us1} = DT_s$
	Detour impedance and home vs. workplace proximity	$T_{us1} = DT_s$ $T_{us2} = R_s$
User-specific attributes		X_{u1}: commuting distance X_{u2}: characteristics of home station X_{u3}: characteristics of workplace station
Activity characteristics		C_{uj1}: activity duration C_{uj2}: activity start time

2.3 Background information and data of the case study

2.3.1 Study area

Shanghai is one of the most populated and fastest growing cities worldwide. The socio-economic development has influenced people's travel behaviour. Local travel surveys show that the trip generation rate of residents has increased in recent years. Meanwhile, the government invested in PT systems to mitigate traffic congestion led by the increasing private car ownership, resulting in an upward trend in the share of PT use (Lu & Gu, 2011). Among all PT modes, the Shanghai metro network is expanding the most in the last years. As shown in Figure 2.2, the metro system operates 14 metro lines, connecting 288 metro stations distributed in the region, among which there are 54 transfer stations (i.e., the stations where passengers can change from one line to another).

A shortest path algorithm can be used to calculate the shortest network distance between every two stations and the number of transfers along each of those paths. The trip fare is set by the operator based on the shortest network distance, and thus they are almost perfectly correlated. The perfect correlation also exists between travel time and network distance, since we assume that the speeds of metro service do not vary between different OD pairs. These are the reasons why in this application we did not use fare and travel time as components of generalized travel costs.

On the website of Dianping[2], which is one of the most popular Chinese location-review services, we mined information of POI, in terms of total number of shops and restaurants within a 500-meter radius from each metro station, indicated by the depth of

[2]http://dianping.com/ (retrieved date: November 21st, 2016)

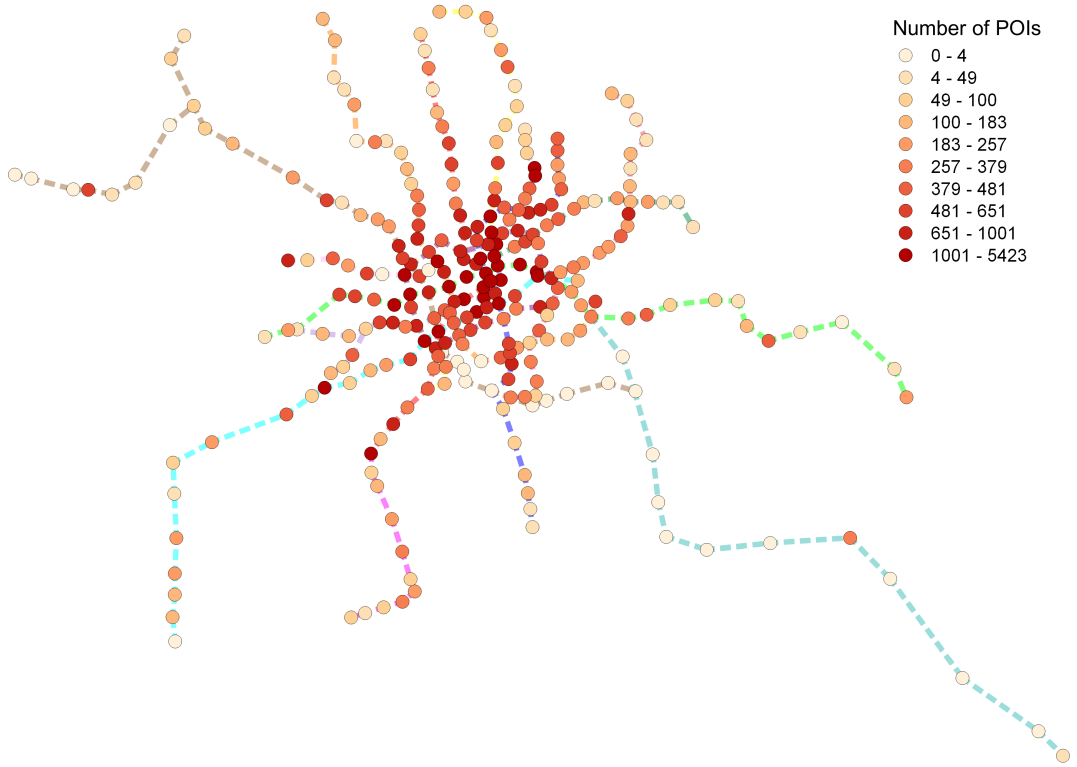


Figure 2.2: The metro network in Shanghai and number of points of interest per station.

colour in Figure 2.2. This variable is regarded as a proxy for the attractiveness of each station for after-work activities in this study. It can be observed that the spatial distribution of POIs is concentrated towards the central part of the city, and it is also interesting to notice that in distant areas from the city centre, that distribution is concentrated in one or two stations, which can be interpreted as being city sub-centres.

2.3.2 Smart card data

One of the ways in which the government promoted PT in Shanghai was to introduce the automated fare collection system that automates the ticketing system for the entire PT network, including metro, bus, taxi, ferry and P+R. Travelers are allowed to pay these services by using a smart card not only for its convenience but also to get a discount.

The SCD provided by the SODA contest contains the records of all transactions by all smart cards in April, 2015. In Shanghai, metro is the only PT system where card holders should both check in and check out. On the other hand, travellers are required to scan their cards only when boarding a bus or alighting a taxi, not to mention that the location information is missing on these modes. Therefore, we focused on the metro network for further analysis and modelling.

In addition, we carefully dealt with those trips including transfers. In Shanghai, only a few metro stations require travellers to check out and then check in again to switch to

another line. Such cases should not be seen as two separate trips. To distinguish them, we used a threshold of 30 minutes between check-out and check-in at those stations. The selection of this threshold is based on the policy by which after 30 minutes without checking in again, the system will regard the next check-in as the start of a new trip. We assume that travellers are aware of this fact, and if they stay at those stations for more than 30 minutes, they must have performed an activity whose utility can compensate for the loss.

2.4 Results of the case study

2.4.1 Detecting metro commuters and extracting daily metro trip chains

After applying the method for detecting the commuters, there were about 0.8 million metro commuters filtered from the data. This number can be compared with the average daily number of unique card IDs scanned for metro trips, which was about 2 million. We did not include those commuters who had detected PT access and/or egress modes such as bus trips connecting with metro trips for commuting. Figure 2.3 shows the spatial distributions of home stations and workplace stations of all the detected metro commuters. By comparing the spatial distributions of home stations, workplace stations and POIs (shown in Figure 2.2 and Figure 2.3), we found that the spatial distribution of home stations was completely different from the ones of workplace stations and POIs, and the latter two were somehow similar to each other.

In our study, we focused on the metro commuters and extracted their daily metro trip chains which only consisted of metro trips. The ten most common types of the daily metro trip chains are plotted in Figure 2.4. Among the metro commuters on an average weekday, about 64.7% performed the home-work-home chain, which was the most common type of trip chains, and at least 13.5% performed the trip chains involving secondary activities. This shows that neglecting this kind of travel patterns may cause the distortion of travel demand prediction.

Among the chain types involving secondary activities, we analyzed the activity start time and activity duration. It was found that Type 10 is more likely to indicate a person who has a lunch break from work, and Type 7 and 9 correspond more to business trips. Type 3, 5 and 8 are more related to the travel patterns of an individual performing an after-work activity.

2.4.2 Model estimation

We focused on the after-work activities which were performed after 16:00 in Chain Type 5. Considering the computational limits, we randomly selected 3,000 commuters

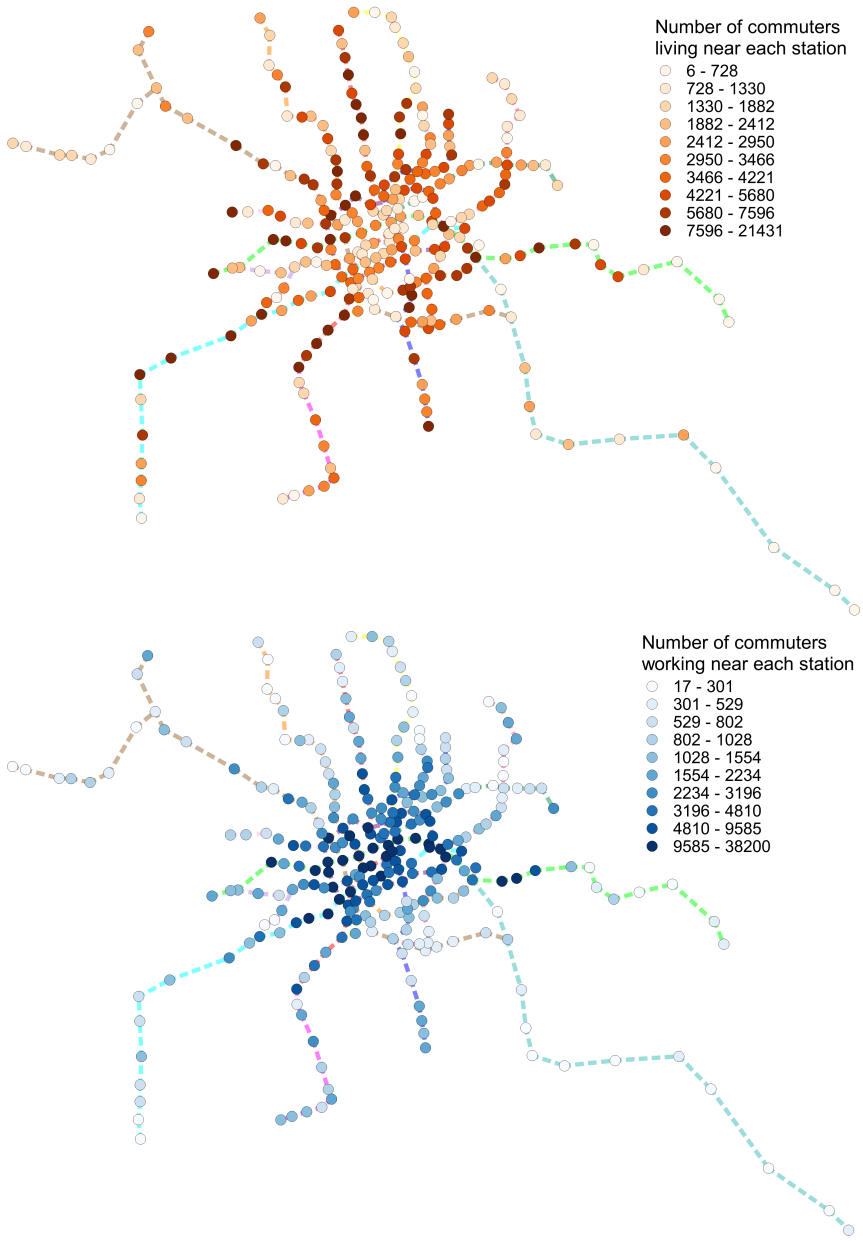


Figure 2.3: Spatial distribution of commuters living and working near each station.

who experienced the prescriptive choice situations in the month. To explain the revealed station choice behaviour, we used the previously proposed model structure. The variable specifications in the utility function formulated as Equation 2.3 should be updated in the context of the case study. The attractiveness of a station for after-work activities was defined as the number of POIs around the station. The features of travel impedance included metro network distance and number of metro transfers. As the characteristics of an after-work activity, activity duration was assumed to be the time gap between the arrival time and the departure time at the station for an after-work activity, and activity start time was quantified by the time gap between 16:00 and the arrival time at the station for the after-work activity.

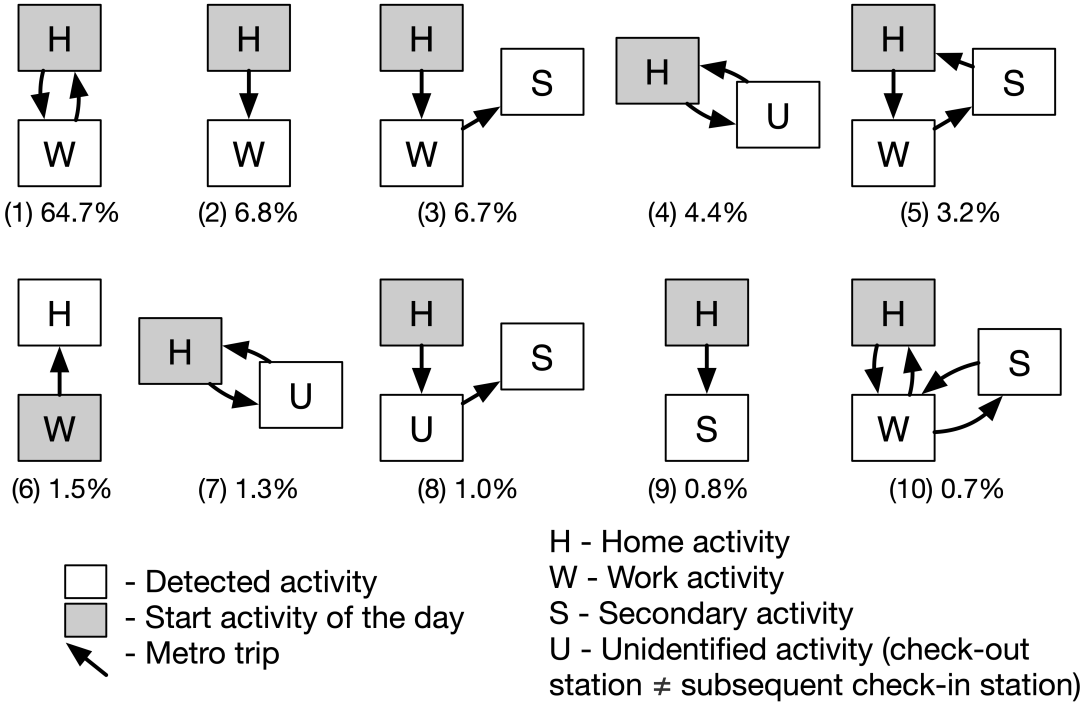


Figure 2.4: The top 10 most common types of daily metro trip chains and their shares.

We have calculated the spatial distribution of home and workplace stations of all the metro commuters (See Figure 2.3). Based on that, we can calculate for each station the ratio of the number of commuters living there over the number of commuters working there, and this ratio is designated as the residents-to-jobs ratio (RJ ratio). For each commuter, we further calculated the RJ ratios for the home station and the workplace station respectively. It can be observed in Figure 2.3 that if the RJ ratio of ones home station is higher, then this person is more likely to live in a mainly residential area, located in the peripheral area of Shanghai; Otherwise, this person is more likely to live in a mainly commercial area, located in the central area of Shanghai. The same applies to interpreting the RJ ratio of ones workplace station. These two variables, along with the commuting network distance and the number of transfers along the commuting trip, can serve as proxies for some personal distinction among the travellers. For each choice occasion, we computed the choice set based on the spatiotemporal constraints. In about 78% of the choice occasions, there is at least one station that a traveller cannot choose due to the constraints.

The estimation results are compared under different model specifications. First, we tested how the different ways of defining travel impedance (See Table 2.1) would influence model fit. Second, we tested how the introduction of the last choice feedback variable would lead to different model estimates.

The estimation results of the models using different kinds of travel impedance without considering last choice feedback are presented in Table 2.4, Table 2.5 and Table 2.6, where only the statistically significant estimates are retained (p-value <0.05). Biogeme (Bierlaire, 2003) is the software package we used for model estimation in this study.

Table 2.4: The estimation results of the discrete choice model using home-based travel impedance without considering last choice feedback.

Variable	**Param.**	**Robust t-test value**
Number of POIs	4.28e-04	10
Number of POIs × activity duration	3.98e-05	6.06
Number of POIs × commuting network distance	3.32e-05	2.43
Number of POIs × RJ ratio of home station	1.53e-05	2.51
Number of POIs × RJ ratio of workplace station	-1.1e-05	-2.27
Network distance from home	-0.255	-12.6
Network distance from home × activity duration	0.014	4.86
Network distance from home × activity start time	-0.00987	-3.19
Network distance from home × commuting network distance	0.0592	9.14
Network distance from home × commuting number of transfers	-0.0223	-4.25
Number of transfers from home	-0.848	-4.62
Number of transfers from home × activity duration	0.184	6.45
Number of transfers from home × commuting network distance	-0.293	-4.89
Number of transfers from home × commuting number of transfers	0.544	9.63
Number of transfers from home × RJ ratio of home station	0.0616	2.38
Number of transfers from home × RJ ratio of workplace station	0.0702	3.29
Number of observations: 5107; Initial log likelihood: -26589.161; Final log likelihood: -21128.36; Adjusted rho-square: 0.205; Run time: 1'58"		

First, the effects of spatial autocorrelation are found to be statistically insignificant in all cases as the estimated values of the dissimilarity parameter are not significantly different from 1. Thus, the spatially correlated model structure actually collapses to the multinomial logit one, of which we present the results. Second, we see that the metro commuters significantly prefer to visit the stations where there are more POIs for performing after-work activities, which is not a surprise. Third, the model using both detour impedance and home vs. workplace proximity fits the data slightly better than the model using only detour impedance, and both of them outperform the one using home-based impedance. This result substantiates the research conclusion drawn by Arentze & Timmermans (2007) regarding the benefit of modelling detour travel impedance, and apart from that, it further shows that commuters do give different weights to travel impedance to access an after-work activity coming from home or from the workplace. It turns out that they generally prefer the stations which are closer from the workplace in terms of number of transfers but closer from home in terms of network distance, ceteris paribus. Fourth, the attributes related to activities are observed to have a considerable impact on station choices for after-work activities. The results significantly show that people give a higher weight to the number of POIs and care less about all kinds of travel impedancs if the activity duration is longer. In addi-

Table 2.5: The estimation results of the discrete choice model using detour travel impedance without considering last choice feedback.

Variable	**Param.**	**Robust t-test value**
Number of POIs	4.44e-04	9.31
Number of POIs × activity duration	6.80e-05	9.13
Number of POIs × RJ ratio of home station	1.59e-05	2.34
Number of POIs × RJ ratio of workplace station	3.03e-05	5.58
Detour network distance	-0.0579	-3.6
Detour network distance × activity duration	0.00914	3.26
Detour network distance × commuting network distance	-0.0174	-3.43
Detour network distance × commuting number of transfers	-0.0134	-2.17
Detour number of transfers	-0.918	-7.58
Detour number of transfers × activity duration	0.180	8.98
Detour number of transfers × commuting network distance	-0.0795	-2.08
Detour number of transfers × RJ ratio of workplace station	0.0840	5.41
Number of observations: 5107; Initial log likelihood: -26589.161; Final log likelihood: -20530.100; Adjusted rho-square: 0.227; Run time: 1'46"		

tion, an activity of longer duration is preferred to take place near the workplace station than near the home station in terms of network distance. The activity start time is an especially effective variable interacting with the home vs. workplace proximity. It can be observed that for a later activity, peoples preference for reducing travel impedance from home weighs more than reducing the one from workplace. Fifth, results seem to support the use of proxy variables to translate differences between travellers. Given that an individual has longer commuting distance, this person seems to be more reluctant to detour farther for after-work activities. A commuter whose home station has higher RJ ratio is more willing to visit a station with a greater number of POIs for after-work activities.

We also estimated the model using detour travel impedance and home vs. workplace proximity after considering last choice feedback. The first choice of each traveller was not modelled since it was assumed to be exogenously given. The estimation results are shown in Table 2.7.

Again the effect of spatial correlation is not statistically significant in this model. It can be observed that travellers frequently chose the same station for after-work activities, leading to the overwhelmingly significant estimate of the preference for the last choice feedback variable which leads to a better model fit. Such a good fit does not necessarily lead to a good demand prediction in future scenarios, because the model relies heavily on the assumption that the previous choice is exogenously given. However, this model can still help us figure out whether we misestimate any parameters due to

Table 2.6: The estimation results of the discrete choice model using detour travel impedance and proximity to home vs. workplace without considering last choice feedback.

Variable	**Param.**	**Robust t-test value**
Number of POIs	4.21e-04	8.73
Number of POIs × activity duration	5.50e-05	7.3
Number of POIs × RJ ratio of home station	2.00e-05	2.94
Number of POIs × RJ ratio of workplace station	2.26e-05	3.98
Detour network distance	-0.0613	-3.84
Detour network distance × activity duration	0.00688	2.47
Detour network distance × commuting network distance	-0.0152	-3.01
Detour network distance × commuting number of transfers	-0.0121	-2
Detour network distance × RJ ratio of home station	0.00589	2.24
Detour number of transfers	-0.931	-7.59
Detour number of transfers × activity duration	0.171	8.47
Detour number of transfers × RJ ratio of workplace station	0.0802	5.1
Home vs. workplace proximity (network distance)	-0.0676	-3.78
Home vs. workplace proximity (network distance) × activity duration	0.0131	5.32
Home vs. workplace proximity (network distance) × activity start time	-0.00717	-3
Home vs. workplace proximity (network distance) × commuting network distance	0.0185	3.22
Home vs. workplace proximity (number of transfers)	0.414	2.74
Home vs. workplace proximity (number of transfers) × activity start time	-0.0863	-3.12
Home vs. workplace proximity (number of transfers) × commuting network distance	-0.113	-2.29
Number of observations: 5107; Initial log likelihood: -26589.161; Final log likelihood: -20404.619; Adjusted rho-square: 0.231; Run time: 3'10"		

neglecting habitual effect. After introducing the variable of last choice feedback, results indicate that travellers actually do not give as much weight to the number of POIs as was estimated previously. To make a choice among those stations which have not been visited previously, people seem to care less about detour number of transfers but care more about detour network distance, and they are more likely to choose a station even closer to home in terms of network distance. The effect of activity start time is no longer significant on the preference for home vs. workplace impedance, indicating that this effect estimated in the previous models might have been related with habitual behaviour. However, the effects of activity duration and commuting network distance

Table 2.7: The estimation results of the discrete choice model using detour travel impedance and proximity to home vs. workplace without considering last choice feedback.

Variable	**Param.**	**Robust t-test value**
Number of POIs	3.88e-04	2.84
Number of POIs × activity duration	7.15e-05	3.28
Number of POIs × commuting network distance	-1.06e-04	-2.81
Detour network distance	-0.0734	-2.84
Detour network distance × activity duration	0.0107	2.61
Detour network distance × commuting network distance	-0.021	-2.94
Detour number of transfers	-0.851	-3.63
Detour number of transfers × activity duration	0.141	4.26
Detour number of transfers × RJ ratio of workplace station	0.0872	3.3
Home vs. workplace proximity (network distance)	-0.144	-6.14
Home vs. workplace proximity (network distance) × activity duration	0.0108	3.22
Home vs. workplace proximity (network distance) × commuting network distance	0.0439	6.24
Home vs. workplace proximity (number of transfers)	0.689	2.24
Home vs. workplace proximity (number of transfers) × commuting network distance	-0.244	-2.69
Last choice feedback	3.99	60.39
Number of observations: 2127; Initial log likelihood: -11448.617; Final log likelihood: -6378.6849; Adjusted rho-square: 0.440; Run time: 2'38"		

on the preferences still exist.

2.5 Conclusions and recommendations

In this paper, after detecting metro commuters and extracting their trip chains from the SCD, we focused on modelling their station choices for after-work activities. The method was applied to the case study of metro travellers in Shanghai. The advantages of using SCD over travel survey data for this purpose include the cost efficiency of data collection, the full population of travellers, and the revealed panel effect. In addition, to overcome the drawback of such anonymous data, we proposed to use proxy variables to distinguish the travellers, which can help better explain the heterogeneity of location choice behaviour among the population. Moreover, different ways of modelling travel impedance were compared, and we found that the model using detour impedance and home vs. workplace proximity, which we created in this study to model the travel

impedance to conduct after-work activities, outperformed the others and improved the interpretation of behaviour.

This work can still be improved in a few ways. First, a travel survey dataset is recommended to be complementarily used for validation and reference. It can help improve the accuracy of commuter detection and identify more specific activity purposes among after-work activities. Also, stated-preference data from travel survey can potentially help enhance the understanding of how travellers perceive travel impedance for after-work activities, further improving our proposed travel impedance metrics. For example, the preference for reducing travel impedance may be related to factors such as familiarity with a particular area, which is difficult to obtain using smart card data. Next, the discrete choice model can be further elaborated to take more factors into consideration. Finally, we only focused on the station choices for after-work activities conducted in a certain type of daily trip chain in this study; however, a more general framework can be built to model station choices for all secondary activities using SCD in future research.

2.6 Acknowledgment

We would like to express our gratitude to the Shanghai Open Data Apps (SODA) contest for making the data available for this research.

Chapter 3

Understanding spatial preferences based on mobile internet usage

---

The previous chapter used socio-geographic status as a proxy variable to segment a given population and further predict travelers' after-work location choice. Today, in this hyper-connected, technological society, one's personal attributes relate not only to one's status in the physical world but also to one's profile on the internet. However, few studies have been conducted to link spatial behavior with mobile internet usage, a gap that this chapter fills. A special dataset from Shanghai, China is used, including individuals' spatial-temporal traces and mobile internet usage, thus revealing the relationship between their preferred types of non-commuting trip destinations and their preferred types of mobile internet content.

The chapter is based on the following publication:

Wang, Y., Correia, G.H.A., van Arem, B., & Timmermans, H.J.P. (2018). Understanding travellers preferences for different types of trip destination based on mobile internet usage data. Transportation Research Part C: Emerging Technologies, 90, 247-259.

Abstract

New mobility data sources like mobile phone traces have been shown to reveal individuals movements in space and time. However, socioeconomic attributes of travellers are missing in those data. Consequently, it is not possible to partition the population and have an in-depth understanding of the socio-demographic factors influencing travel behaviour. Aiming at filling this gap, we use mobile internet usage behaviour, including ones preferred type of website and application (app) visited through mobile internet as well as the level of usage frequency, as a distinguishing element between different population segments. We compare the travel behaviour of each segment in terms of the preference for types of trip destinations. The point of interest (POI) data are used to cluster grid cells of a city according to the main function of a grid cell, serving as a reference to determine the type of trip destination. The method is tested for the city of Shanghai, China, by using a special mobile phone dataset that includes not only the spatial-temporal traces but also the mobile internet usage behaviour of the same users. We identify statistically significant relationships between a travellers favourite category of mobile internet content and more frequent types of trip destinations that he/she visits. For example, compared to others, people whose favourite type of app is in the tourism category significantly preferred to visit touristy areas. Moreover, users with different levels of internet usage intensity show different preferences for types of destinations as well. We found that people who used mobile internet more intensively were more likely to visit more commercial areas, and people who used it less preferred to have activities in predominantly residential areas.

Keywords: Mobile internet usage; mobile phone data; travel behaviour; mobility analysis; data fusion.

3.1 Introduction

There is a recent trend in complementing or even replacing traditional travel survey data with new mobility-related data sources, such as GPS data, mobile phone traces and smart card transaction data (Chen et al., 2016; Demissie et al., 2013b; Iqbal et al., 2014; Ni et al., 2018; Toole et al., 2015; Wang et al., 2017; Wolf, 2006; Yue et al., 2014; Zhao et al., 2018). These trajectory-based data are getting popular for travel analysis because (1) they are inexpensive to collect; (2) they are usually up to date; and (3) most of them contain a large sample with observations that are longitudinal in time (Calabrese et al., 2013; Demissie et al., 2013a; Morency et al., 2007).

However, despite the potential advantages, these sources of information only include the spatial-temporal traces describing peoples movements. If the aim is to understand travel behaviour from an activity-based perspective (Chen et al., 2016; Rasouli & Timmermans, 2014; Zhao & Zhang, 2017), the information of these data sets is usually very limited. For example, activity purpose of the trips is typically missing (Calabrese

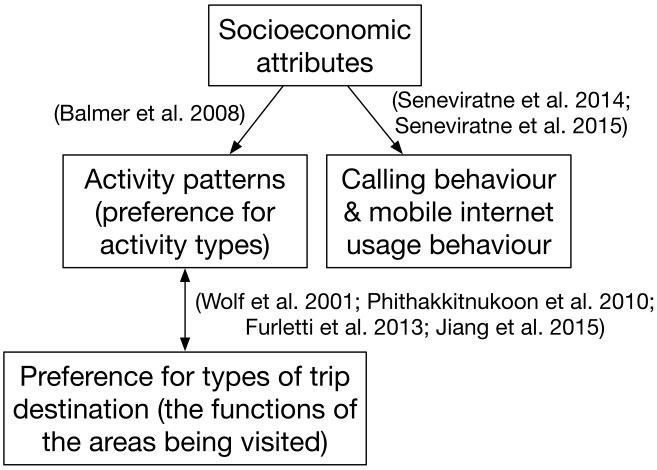


Figure 3.1: The conceptual framework.

et al., 2013). Moreover, in traditional travel demand models, socioeconomic information is used to segment the population, and better explain the heterogeneity of activity-travel behaviour, including, but not limited to, activity patterns (Balmer et al., 2008) and location choice (Sivakumar & Bhat, 2007). However, in anonymous big data, socioeconomic information is unavailable mainly due to privacy reasons (Calabrese et al., 2015).

To deal with such problems, researchers have tried to combine different types of data in order to fill the gaps (Anda et al., 2017). In attempting to derive activity purpose information from trajectory data, there have been several applications fusing trajectory data with land use data, OpenStreetMap data or point of interest (POI) data (Dashdorj et al., 2013; Demissie et al., 2015; Wolf et al., 2004; Yuan et al., 2012). This geo-coded background knowledge can help estimating the function of an area, which can tentatively be connected to the type of activity that a visitor performed in that area (Furletti et al., 2013; Jiang et al., 2015; Phithakkitnukoon et al., 2010; Wolf et al., 2001). We referred to the main function of an area being visited as type of trip destination in this paper. The left chain in Figure 3.1 shows how we derive the dependency of ones preference for destination types on socioeconomic attributes, based on literature review. Intuitively, such dependency exists in most cities. For example, it is common that some specific urban areas are more frequented by young people.

To partition the population using mobile phone data, Arai et al. (2014) and Bwambale et al. (2017) suggested utilizing calling behaviour such as calling frequency and duration to predict ones personal attributes. However, mobile phones are less used for calls today, making calling behaviour less useful, while simultaneously people are spending more time on services provided by mobile internet such as mobile apps (Richmond, 2012). Therefore, mobile internet usage behaviour, if available, could have a greater potential to reflect individuals traits, such as gender and age (Seneviratne et al., 2015, 2014). The right chain in Figure 3.1 shows the dependency of mobile internet usage behaviour on socioeconomic attributes.

As a whole, Figure 3.1, which can be regarded as a conceptual framework, shows the relationship between mobile internet usage behaviour and preference for types of trip destination. Since they are both dependent on the socioeconomic attributes, even if the socioeconomic attributes are unobserved, they are still likely to be correlated with each other. Based on this hypothesis derived from the conceptual framework, our study aims to understand travellers preferences for types of trip destination by means of segmenting them based on the preferred type of sites and applications visited through mobile internet as well as the level of visiting frequency, by fusing mobile phone traces and mobile internet usage data. We are allowed to do this study because of the data provided by the Shanghai Unicom WO+ Open Data Application Contest[1].

Furthermore, mobile internet usage behaviour might sometimes be able to reflect even more information about a person, such as ones specific interests and lifestyles, than the traditional socioeconomic attributes do. At the same time, ones interests and lifestyles are regarded as the determinants of location choice through preference for different types of non-work activities (Wen & Koppelman, 2000). A more specific interest or lifestyle might be related to a more specific travel preference especially for non-work activities. For example, a foodie would visit more sites and applications about food, and meanwhile, he/she would also like to visit more restaurants in real life. We see the potential to explore such relationships by fusing mobile internet usage data and mobile phone traces, and we especially focus on the types of destinations for out-of-home non-work activities, designated herein as secondary activities for simplicity. Many studies have used mobile phone data to analyse users home and workplace locations as well as commuting trips (Ahas et al., 2010; Alexander et al., 2015; Calabrese et al., 2011; Isaacman et al., 2011). However, trips for secondary activities have not often been analysed using this type of data, except in only a few studies (e.g. Huang & Levinson, 2015; Järv et al., 2014), which does not mean that they are not an important part of urban travel demand. In fact, they are taking a larger share than ever before, especially in large metropolitan areas (Wang et al., 2017).

The rest of this paper is organized as follows. First, we introduce the data used in our research. Next, we explain our research method. Then, the results are presented. Finally, we draw the conclusions, discuss the usefulness and limitation of our research, and point out the directions for future research.

3.2 Case study

In this paper, the case study is conducted in Shanghai, China. As one of the four directly-controlled municipalities of China, Shanghai is world famous for being a global financial centre and transport hub. The total area of Shanghai is 6,340 square kilometres, and the population of Shanghai has exceeded 24 million. The city of Shanghai is divided into 16 districts. Except the Chongming district composed of

[1] https://www.kesci.com/woplus/

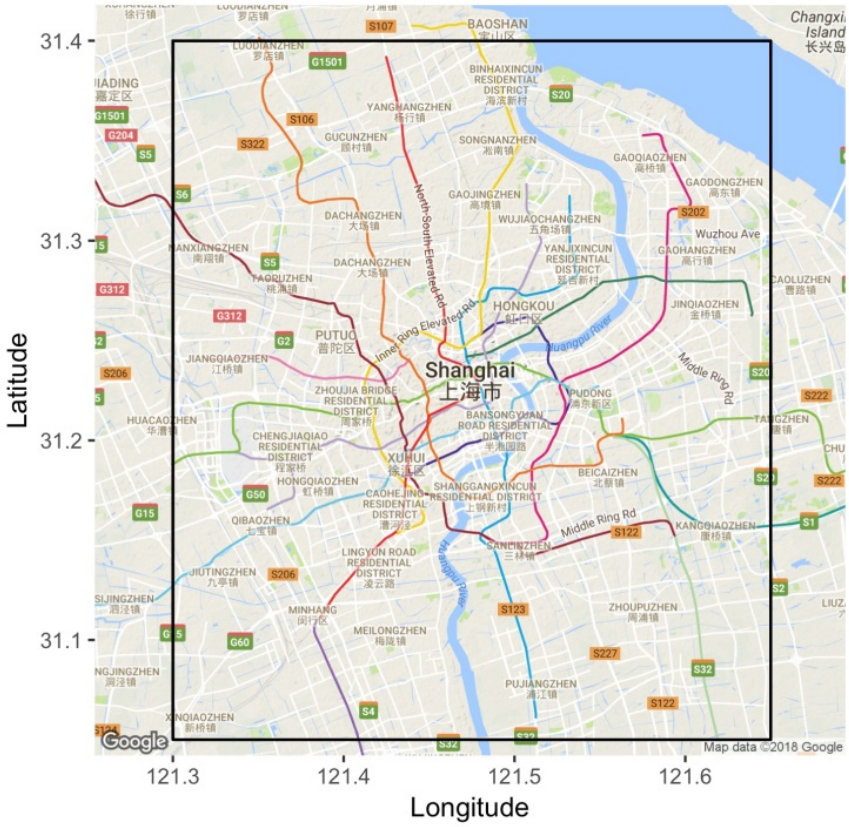


Figure 3.2: The map of the target area.

three islands in the Yellow Sea, the other 15 districts lie on Chinas east coast. They are separated by Huangpu River into two banks, Pudong and Puxi, which literally mean the east bank and the west bank respectively in Chinese. Despite the crossing river, these two banks are well connected by several bridges, tunnels for cars and tunnels for metro. The boundary of our target area, covering the relatively more populous parts of Shanghai, is represented by the rectangle (about 1,775 square kilometres) in Figure 3.2, using the WGS 84 (EPSG:4326) reference coordinate system. Note that WGS 84 / Pseudo-Mercator (EPSG:3857) is used as the projection system to calculate distance in this work.

3.2.1 Mobile phone data

The Shanghai Unicom WO+ Open Data Application Contest provides both the mobile phone traces and the mobile internet usage data of the same sample of the Shanghai Unicom users. Unicom is one of the three mobile carriers in China. It was reported that the total number of the Unicom mobile users had reached about 270 million in China by the beginning of 2017.

The mobile phone traces include the spatial-temporal records of 620 thousand sampled users moving within the city of Shanghai hour by hour from 12 a.m. 27th of December

2015, to 3 p.m. 6th of January 2016. Every time a user had a mobile phone activity (i.e., a call, a text message, a voice mail, or an internet connection), the location information and the timestamp of the activity would automatically be recorded with the anonymous user ID in the original database. However, the provided data were hourly aggregated for each user. To be specific, if a user was detected to have visited several locations within an hour, only the location where the user stayed for the longest time would be known for that hour. It is also possible that a user did not have any mobile phone activity within an hour, thus leaving no location information. It is regarded as a missing trace for that user. The detected location information of an available trace is represented by a pair of coordinates using the WGS 84 (EPSG:4326) reference coordinate system. 4 digits of the longitude coordinate are stored after the decimal point, and 5 digits of the latitude coordinate are stored after the decimal point. According to the data provider, due to the inherent detection inaccuracy, the real location of a trace lies within the 200 m 200 m square of which the centre is the detected point.

The mobile internet usage data include the page view counts of each user for different types of mobile apps and websites during the same study period. The page view counts of mobile apps and websites were merged for the same category, thus producing a total of 13 types of mobile internet contents: finance, food, shopping, social news, housing, tourism, sports, car, entertainment, education, job seeking, game, and health. The specific mobile apps and websites in each category were selected by the data provider. The users who never browsed any mobile internet contents are labelled with the tag null.

3.2.2 POI data

A POI is a specific point location associated with a pair of coordinates and some information about this location, such as name, category and description. The POI data used in our study were extracted from the Gaode Maps service[2], which is the Chinese equivalent of Google Maps. The Gaode open platform allows the registered developers to obtain the POI data of a specific area through the application program interface (API). In our target area, about 260 thousand POIs of ten predefined categories can be obtained. The available information of the POI data includes name, coordinates and category. The ten categories are hotel, sports and recreation, finance and insurance, residence, education, workplace, restaurant, car service, tourism, and health.

3.3 Methodology

In Figure 3.3, we present a flowchart of the proposed research method in this study. First, trip destinations chosen by the users for secondary activities can be extracted

[2]https://lbs.amap.com

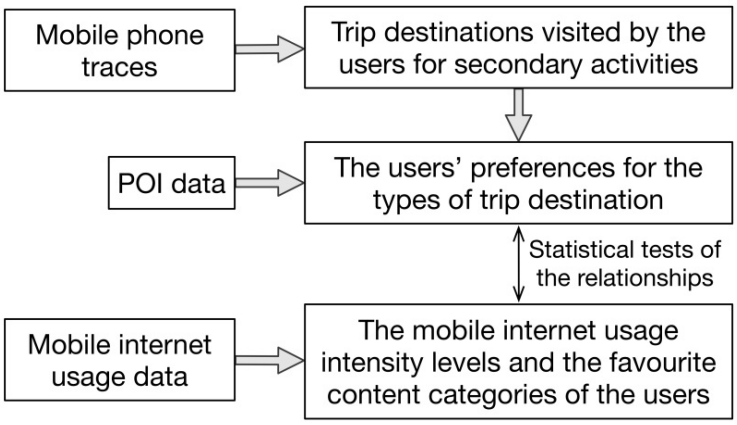


Figure 3.3: The flowchart of the research method.

from mobile phone traces. Second, each trip destination can be labelled by the cluster of the grid cell calculated based on the POI data, and we can discover the users preferences regarding the types of trip destinations for secondary activities. Third, the favourite categories of mobile internet contents and the total usage intensity levels of the users can be derived from mobile internet usage data. Fourth, the relationships can be statistically tested between the users mobile internet behaviour and the users preferences for the types of trip destinations. We also perform a sensitivity analysis to examine to what extent results would be affected due to the inherent spatial inaccuracy of mobile phone traces.

3.3.1 Extracting trip information from mobile phone traces

To ensure that the problem of missing traces would not affect our analysis, we first filter the users and only focus on those who were traced at least 80% of the total hours. Also, we only focus on the users who were always traced within the prescribed boundary of the target area. To estimate the trips made by mobile users, it is important to distinguish stay locations (i.e., origins and destinations of trips) from pass-by locations, in the mobile phone traces (Ahas et al., 2010; Alexander et al., 2015; Wang & Chen, 2018; Zheng et al., 2009). Meanwhile, signal errors may lead to false movement of traces which do not represent actual movement of users (Çolak et al., 2015). The effects of such errors should be reduced as well. In this study, we adopt the main steps lately suggested by Alexander et al. (2015) to detect stay locations, whilst the parameter used in the third step is modified to suit our case:

- For each user, we find the traces that are spatially close (within 300 metres) to their subsequent observations and thus obtain the sets of geographically and temporally close traces. The medoid of the coordinates within each set is then calculated to update the locations of the traces.

- The traces that are close in space but far apart in time need to be consolidated for each user as well. The complete-linkage hierarchical clustering algorithm is applied using 500 metres as the threshold. In this algorithm, we first treat each point as a cluster and merge step by step the two clusters whose merger has the smallest diameter, until the smallest diameter reaches 500 metres. The medoid of the coordinates within of each cluster is used to update the location of the traces.

- To identify whether a user stays or passes through, a duration threshold should be chosen dependent on the assumed shortest activity duration as well as the sampling rate of the data. In our study, the sampling rate for each user is relatively small: at most one trace per hour. Hence, it is stipulated that at least two consecutive traces close in space can determine a stay point, which will necessarily lead to overlooking some short activities; however, this is the best that can be done to extract stay points with these data.

In this research, we will not study the trips toward home or work activities but focus on the trips for secondary activities. Thus, we need to detect them using the stay traces of the users. A possible way is to infer activity purposes based on the ground truth (i.e., the features related to a certain activity purpose). Those features can be sourced from either general knowledge, such as the fact that people mostly spend their night at home (Alexander et al., 2015; Nanni et al., 2013), or travel survey data, which provide more powerful evidence (Liu et al., 2013). In this study, since travel survey data are not available, we apply the rules suggested by Alexander et al. (2015) to detect the trips for secondary activities, and we choose the parameters that suit our case:

- For each user, the home location is defined as the location with most stay traces from 7 p.m. to 8 a.m. on weekdays, on weekends, and on holidays.

- The work location is defined as the place to which one travels the maximum accumulated distance from home, $max(v \times d)$, where v is the number of visits between 8 a.m. and 7 p.m. on weekdays during the study period, and d is the distance between a given place and home location. If the user visits the detected work location fewer than 2 days per week, it is not regarded as a work location.

- It is assumed that the stay traces at the detected home location should be labelled as home activity. The same applies to labelling work activity, and the remaining stay traces are labelled as secondary activity. In this suggested approach, only stable home and workplace locations can be detected.

In the further analysis, we only focus on the users who had a home location and performed at least one secondary activity during the study period.

3.3.2 Clustering types of trip destinations for secondary activities

In this section, we further distinguish the trips for secondary activities in terms of the types of trip destinations. In traditional travel demand models, the purposes of secondary activities, such as eating out and shopping, can be used to distinguish the trips for secondary activities. However, it is very difficult to detect such purposes in mobile phone traces, especially without any travel survey data available as a reference. A compromising solution is to distinguish the trips for secondary activity based on geographical information of the visited area, such as land use (Wolf, 2006), and the POI data can be used to depict urban land use in a more detailed way (Jiang et al., 2015; Phithakkitnukoon et al., 2010; Yuan et al., 2012). Following this strategy, we define the types of trip destinations for secondary activities as follows.

A virtual grid reference can be constructed to divide the city (Demissie et al., 2015; Phithakkitnukoon et al., 2010). Each cell should be characterized and serves as a reference for determining the type of trip destination. For a cell $k \in \{1,2,...,K\}$, the number of each type of POIs is calculated, named p_{kj}, where $j \in \{1,2,...,J\}$ indicates a POI type (e.g., restaurant or workplace). The number of POIs of each type is then ranked over all cells, and the percentile rank r_{kj} is calculated as the percentages of cells that have lower number of POIs of type j than cell k has. As a result, each cell k can be portrayed as a vector of the percentile ranks of all the POI types $\boldsymbol{r}_k = (r_{k1}, r_{k2}, ..., r_{kJ})$.

In our study, a hierarchical clustering algorithm is applied to the vectors of all cells. We use the Pearson-correlation-based distance metric (Resnick et al., 1994; Xue et al., 2005) since we assume that the similarity between the functions of two areas can be reflected by the correlation between vector $\boldsymbol{r}_k$ and vector $\boldsymbol{r}'_k$, where $k' \in \{1,2,...,K\}\ \{k\}$. The distance $d_{kk'}$ between these two vectors, used in the clustering algorithm, is calculated in the following equation:

$$d_{kk'} = 1 - \frac{cov(\boldsymbol{r}_k, \boldsymbol{r}'_k)}{s(\boldsymbol{r}_k)s(\boldsymbol{r}'_k)} \tag{3.1}$$

where $cov(\boldsymbol{r}_k, \boldsymbol{r}'_k)$ is the covariance of $\boldsymbol{r}_k$ and $\boldsymbol{r}'_k$; $s(\boldsymbol{r}_k)$ is the standard deviation of $\boldsymbol{r}_k$; $s(\boldsymbol{r}'_k)$ is the standard deviation of $\boldsymbol{r}'_k$. Since correlation is scale-invariant, it is better to standardize $\boldsymbol{r}_k$ as $\hat{\boldsymbol{r}_k}$ to represent the profile of a cell, whose element is calculated as follows:

$$\hat{r_{kj}} = (r_{kj} - \bar{\boldsymbol{r}_k}) / \sqrt{\sum_j (r_{kj} - \bar{\boldsymbol{r}_k})^2} \tag{3.2}$$

Where $\bar{\boldsymbol{r}_k}$ is the mean of $\boldsymbol{r}_k$. To find relatively more compact clusters of approximately equal diameters, we choose the complete-linkage clustering method (Everitt et al.,

2011). Consequently, each cell k can be related to a cluster $c \in \{1,2,...,C\}$. Cluster compactness can be assessed by the Dunn index (Dunn, 1973), which is the ratio of the smallest distance between observations in different clusters to the largest intra-cluster distance. Note that the distance used to calculate the Dunn index is still the Pearson-correlation-based distance defined previously. Intuitively, maximizing the Dunn index can help us select the optimal parameters and obtain the most distinctive urban area functions. In this case, the parameters to be selected include the number of clusters and the side length of the grid cells.

We pre-set the upper bound of the number of clusters as 10, equal to the number of dimensions of the POI data in this study, mainly for interpretation. In this study, we aim to interpret the statistical relationship between the preference for mobile internet contents and the preference for types of trip destination. Each type of trip destination is desired to have a distinctive characteristic. Thus, we expect our clusters to reflect the most distinctive urban functions. If the number of clusters is too large, the differences between some urban functions would possibly become very subtle, and the corresponding types of trip destination would be difficult to interpret.

Different from the other studies choosing an arbitrary value for the side length of the grid cells of the city, for example, 500 meters (Phithakkitnukoon et al., 2010) and 800 meters (Demissie et al., 2015), our study tests several values (i.e., 300 meters, 400 meters, 500 meters, 600 meters, 700 meters and 800 meters) as the side length of a grid cell. We only consider this range of values because the size of the grid cell should neither be too large nor too small. If it is too large, the defined function of a cell must become too rough; if it is too small, the detected destination of a trip would be very likely to lie in a wrong cell due to the inherent detection inaccuracy explained in Section 3.2.1. However, even if the size of a grid cell is very large, it is still possible that the detected destination and the real destination would lie in different grid cells. Thus, we will present the method to examine the impact of this issue on the final results in Section 3.3.5.

We generate the clustering results iteratively and choose the combination of the side length value and the number of clusters that can maximize the Dunn index and thus give us the most compact set of clusters. Consequently, each user has a set of trips for secondary activities during the study period. The coordinates of a trip destination can correspond to a grid cell k and further correspond to a cluster c, which is defined to be the type of that trip destination.

3.3.3 Analysing mobile internet usage behaviour

Let f_{un} indicate the frequency of browsing a type of mobile internet content $n \in \{1,2,...,N\}$ (e.g., finance or shopping) through mobile apps and/or websites by an individual $u \in \{1,2,...,U\}$ across several days. Given this, two main indicators of ones mobile internet usage during a period can be derived: (1) the frequency of using

all mobile internet contents $F_u = \sum_n f_{un}$, which reflects an individuals usage intensity, and (2) the relative preferences for using different types of mobile internet contents, expressed in terms of an N-dimensional vector $w_u = (w_{u1}, w_{u2}, ..., w_{uN})$, reflecting the different lifestyles and interests. We rank f_{un} for each n over all users and calculate w_{un} as the percentages of users who browse n less often than user u does.

Based on the total usage intensity, the population can be divided into three classes: (1) the null class, representing the people who never use any mobile internet service, (2) the low intensity class, representing the people whose usage intensity is lower than or equal to the median value of all non-zero total usage intensities, and (3) the high intensity class, representing the people whose usage intensity exceeds the median value of all non-zero total usage intensities. To segment the population using the preferences for specific contents, we find the content category n that maximizes w_{un} for a user u and use it to tag this user. Intuitively, such a tag is a users favourite content category. For example, a user can predominantly be tagged as shopping, finance, etc.

3.3.4 Relating preferred types of trip destinations to mobile internet usage behaviour

In order to understand if there are statistically significant differences regarding the preferences for the different types of trip destinations among those who have different preferences for mobile internet content, we mainly use the statistical test of comparing two population proportions with independent samples, which is explained as follows.

The number of trips going to a destination of type c is aggregated over the users tagged as n regarding mobile internet content. The aggregate number of these trips is expressed as x_{cn}, and the total number of trips made by the users with the interest tag n is $x_n = \sum_c x_{cn}$. Then the proportion of trips to the destinations of type c made by the users with the interest tag n is $\rho_{cn} = x_{cn}/x_n$. On the other hand, the number of trips to the destinations of type c is aggregated over the remaining users who do not prefer n. The aggregate number of these trips is expressed as $x_{cn'}$, where $n' \in \{1, 2, ..., N\} \setminus \{n\}$, and the total number of the trips made by the remaining users is $x_n = \sum_c x_{cn'}$. The proportion of trips to the destinations of type c made by the remaining users is $\rho_{cn'} = x_{cn'}/x_{n'}$. The two-tailed z-test, if following a normal distribution, is appropriate for our objective, which is to check whether the two proportions, ρ_{cn} and $\rho_{cn'}$, are different or the same, and the test statistic is given as follows:

$$Z = (\rho_{cn} - \rho_{cn'}) / \sqrt{\rho_{cn}{}^*(1 - \rho_{cn}{}^*)(1/x_n + 1/x_{n'})} \tag{3.3}$$

where $\rho_{cn}{}^* = (x_{cn} + x_{cn'})/(x_n + x_{n'})$.

Based on the value of Z, the significance of the difference can be derived, in terms of the corresponding p-value pv_{cn}. For every combination of c and n, we calculate the

significance of the difference. Thus, there are $C \times N$ cases in total, causing the multiple comparisons problem: if a statistical analysis involves multiple simultaneous statistical tests, there will be more chances of rare events, increasing the likelihood of incorrectly rejecting a null hypothesis (Rupert Jr et al., 2012). Therefore, a stricter threshold of p-value should be used to reject a null hypothesis. In this study, we use the Bonferroni correction, which suggests dividing the original p-value threshold by the number of hypotheses. In our case, we set the original p-value threshold as the typical one, 0.05, and the threshold after the Bonferroni correction is $0.05/(C \times N)$.

We construct an indicator of the significance of the preference $pref_{cn}$, explained in the following equation:

$$pref_{cn} = \begin{cases} log_{10}(1/pv_{cn}) & \text{if } \rho_{cn} \geq \rho_{cn'} \text{ and } pv_{cn} < 0.05/(C \times N) \\ -log_{10}(1/pv_{cn}) & \text{if } \rho_{cn} < \rho_{cn'} \text{ and } pv_{cn} < 0.05/(C \times N) \\ 0 & \text{if } pv_{cn} \geq 0.05/(C \times N) \end{cases} \tag{3.4}$$

The absolute value of this indicator is larger if the significance is higher. If the indicator is positive, it means that compared to the others, the users tagged by n significantly prefer to visit the destinations of type c. If the indicator is negative, it means that the users tagged by n significantly prefer not to visit the destinations of type c. If the indicator is zero, it means that there is no significantly different preference.

The same method can be applied to understand if there are statistically significant differences regarding peoples preferences for different types of trip destinations among those who have a certain level of total mobile internet usage intensity.

3.3.5 Sensitivity analysis

Consider T_{ui} as the ith stay trace of an individual $u \in \{1,2,...,U\}$. The location of a stay trace can be represented by longitude lon_{ui} and latitude lat_{ui} in terms of metres using the WGS 84 / Pseudo-Mercator (EPSG:3857) projection system. As mentioned in Section 3.2.1, the true location of a trace lies within the $200m \times 200m$ square of which the centre is the detected point. Thus, it is possible that the true activity location does not lie in the correct grid cell. In this study, we assess the impact of such detection inaccuracy on the results regarding the statistical relationship between the preference for mobile internet contents and the preference for types of trip destination.

We assume that the longitude of the true location of a stay trace lon_{ui} can be uniformly drawn inside the interval $[lon_{ui} - 100, lon_{ui} + 100]$, and the latitude of the true location lat'_{ui} can be uniformly drawn inside the interval $[lat_{ui} - 100, lat_{ui} + 100]$. We draw lon'_{ui} and lat'_{ui} of all the stay traces independently in 20 loops, except in the first loop where we set lon'_{ui} as lon_{ui} and set lat'_{ui} as lat_{ui}. In each loop, based on lon'_{ui} and lat'_{ui}, the stay traces are assigned to their belonging grid cells. Consequently, the type of the trip

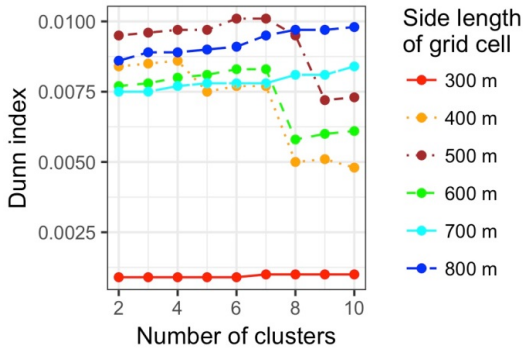


Figure 3.4: The Dunn index used to determine the number of clusters and the side length of the grid cells.

destination corresponding to each detected trip can be determined in each loop. Finally, we mainly assess the two specific impacts on the results of the statistical relationships. First, we assess whether any conflicting significant results will be found in the 20 loops. Second, we examine whether the same significant relationships are robust enough to be found in more than 80% of the loops, namely 16 loops. We construct an indicator of the significance of the robust preference $pref'_{cn}$, explained in the following equation:

$$pref'_{cn} = \begin{cases} log_{10}(1/\overline{pv_{cn}}) & \text{if } \rho_{cn} \geq \rho_{cn'} \text{ and } pv_{cn} < 0.05/(C \times N) \text{ in} \geq 16 \text{ rounds} \\ -log_{10}(1/\overline{pv_{cn}}) & \text{if } \rho_{cn} < \rho_{cn'} \text{ and } pv_{cn} < 0.05/(C \times N) \text{ in} \geq 16 \text{ rounds} \\ 0 & \text{Otherwise} \end{cases} \tag{3.5}$$

Where $\overline{pv_{cn}}$ is the average of the values of pv_{cn} in the loops, and $pv_{cn} < 0.05/(C \times N)$. The same method can also be applied to construct an indicator of the robust statistical relationships between mobile internet usage intensity and preferred types of trip destination.

3.4 Results and discussion

We processed the mobile phone traces using the method explained in Section 3.3.1. As a result, we obtained 26,535 target users meeting the specified criteria, and we detected their trips for secondary activities. Next, we clustered the grid cells using the method explained in Section 3.3.2. As shown in Figure 3.4, based on the Dunn index, we can find that the clusters are best distinguished by setting the number of clusters as 6 or 7 and setting the side length as 500 meters in our case. We chose the smaller number of clusters, 6, for interpretation.

Figure 3.5 geographically shows the computed clusters of the grid cells in the city. Table 3.1 and Figure 3.6 show the portrait of each cluster, where the profile chart

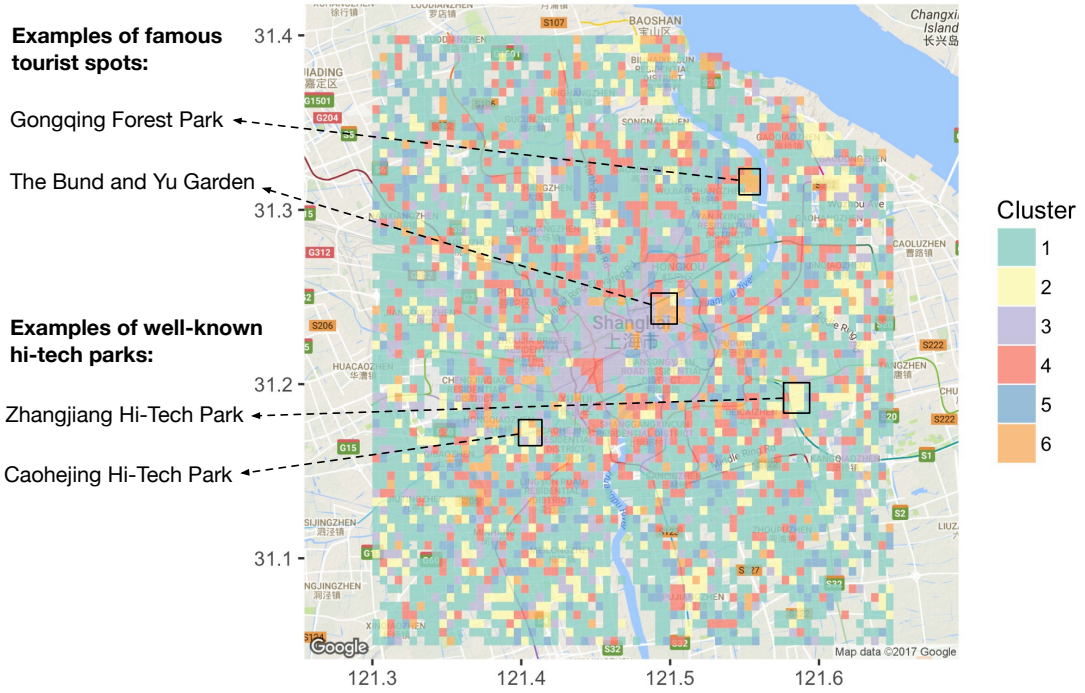


Figure 3.5: The clustered grid cells of the city.

depicts the first quartile, the median and the third quartile of $\hat{r_{kj}}$ (see Section 3.3.2 for the definition) of the cells belonging to each cluster. $\hat{r_{kj}}$ can indicate which POI types are dominant in a cell. For a cell k, the POI type j is relatively more influential if the value of $\hat{r_{kj}}$ is higher.

In Table 3.1 and Figure 3.6, it appears that the variances of $\hat{r_{kj}}$ for restaurant, education and sports & recreation are relatively small among the clusters; hence, they are not the main factors making a cluster different from the others. On the other hand, the other POI types all seem to determine the characteristics of a cluster. Among them, the POI types of residence, workplace and tourism play the most important role in distinguishing the clusters. It can be observed that cluster 2, 3 and 6 represent the areas with relatively more workplaces, whilst cluster 1, 4 and 5 represent the areas with relatively fewer workplaces. Thus, it is not a surprise to see that the relative importance of residence is higher in cluster 1, 4 and 5. Among those more commercial clusters, cluster 3 is a special one since it seems to be all-round in terms of the relatively higher importance of most POI types including residence. Moreover, the relative importance of tourism is nearly zero in cluster 1, 2 and 5; in contrast, it is the highest in cluster 6, which thus seems to represent predominantly touristy areas.

It can be observed in Figure 3.5 that the city centre is mostly composed of the cells belonging to cluster 3, implying that the centre is more multifunctional compared to the remaining parts. In addition, we can find in Figure 3.5 that the famous tourist spots such as Gongqing Forest Park, Yu Garden and the Bund all belong to cluster 6, and the well-known hi-tech parks such as Caohejing and Zhangjiang are assigned to cluster 2. These results make much sense based on our interpretation of the cluster profiles.

Table 3.1: The portraits of the six clusters.

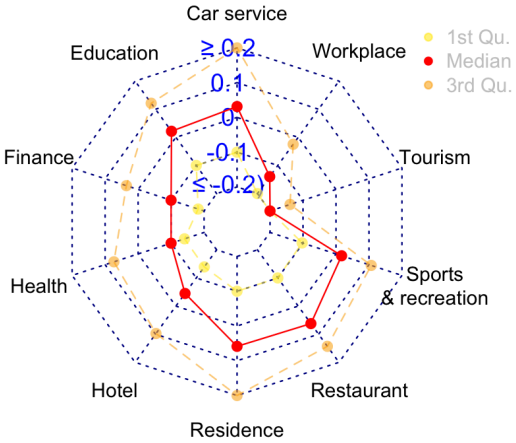
Cluster	Description		Profile chart (in terms of $\hat{r_{kj}}$)
	Based on the profile	**Based on the location**	
1	Residential (with more car service POIs)	Mainly outside the centre	Subfigure 3.6a
2	Commercial (or industrial)	Mainly outside the centre	Subfigure 3.6b
3	All-round	Mainly in the centre	Subfigure 3.6c
4	Residential (more multifunctional)	Mainly outside the centre	Subfigure 3.6d
5	Residential (with more health POIs)	Mainly outside the centre	Subfigure 3.6e
6	Touristy and commercial	Some in the centre and some outside	Subfigure 3.6f

We also characterized each cluster with a few keywords in Table 3.1. Note that although cluster 1, 4 and 5 all somehow represent the predominant residential areas mainly outside the city centre, they are still different in terms of the relative importance of the other types of POIs within each cluster. In the areas belonging to cluster 1, there are relatively more POIs of car service in addition to residence. On the other hand, there are relatively more POIs of health in the areas belonging to cluster 5. The areas belonging to cluster 4 seem more multifunctional. They are almost similar with the areas belonging to cluster 3, except that they have a very low number of workplaces.

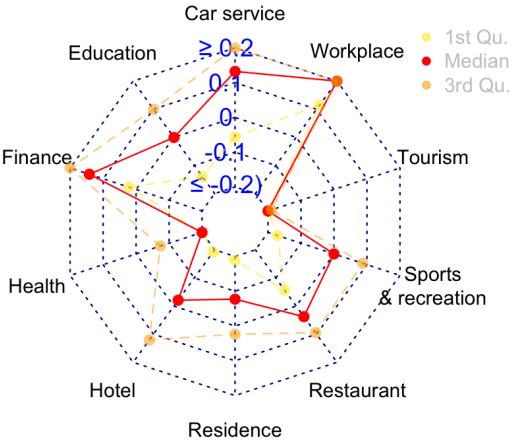
Figure 3.7 presents the results of the statistical test between the mobile internet usage behaviour and the preferences for the types of trip destinations in the initial loop using the original spatial traces (see the explanation in Section 3.3.5). Intuitively, if a category of users like/dislike visiting the destinations of a specific type more significantly, the corresponding colour, representing the indicator $pref_{cn}$ (see the explanation in Section 3.3.4), will be deeper red/blue. Based on our definition of $pref_{cn}$ (equal to zero for the insignificant results), only the significant results are retained in the figure.

Next, we randomly draw the locations of the traces within the boundaries in 20 loops to examine the impact of mobile detection inaccuracy. We first found that there were no conflicting statistical relationships (e.g., ρ_{cn} is significantly larger than $\rho_{cn'}$ in one loop, but significantly smaller than $\rho_{cn'}$ in another loop) in these 20 loops. Figure 3.8 further presents the results of the robust statistical relationships that held in more than 16 loops. Comparing Figure 3.7 and Figure 3.8, we can find that the preferences of people tagged by health, shopping and social news for certain types of trip destination did not hold in more than 16 loops, and the preferences of people who did not use any mobile internet were also sensitive to the possible spatial detection errors.

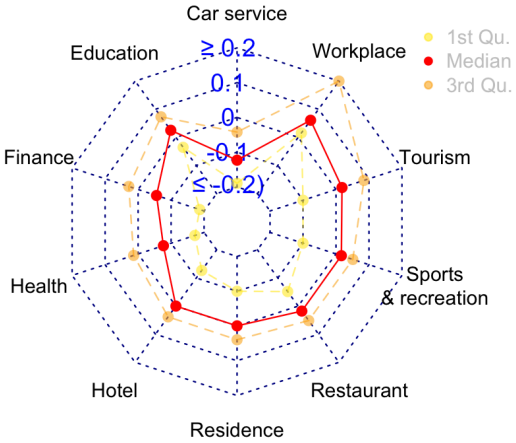
Results show that people who have different tastes in mobile internet content do have



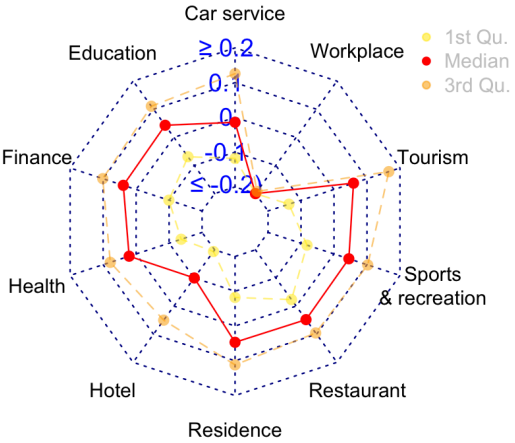
(a) The profile chart of Cluster 1.



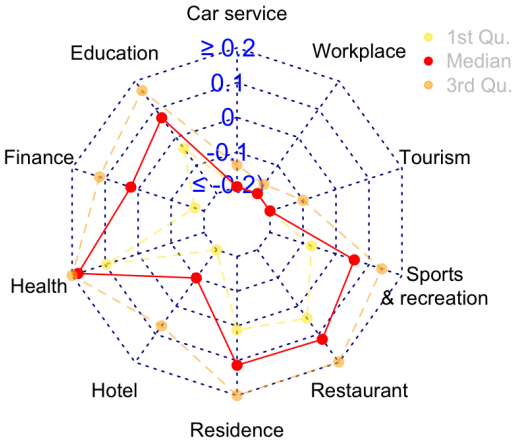
(b) The profile chart of Cluster 2.



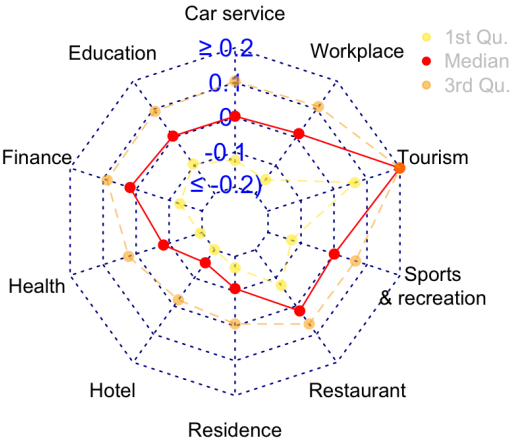
(c) The profile chart of Cluster 3.



(d) The profile chart of Cluster 4.



(e) The profile chart of Cluster 5.



(f) The profile chart of Cluster 6.

Figure 3.6: The profile charts of the six clusters.

different preferences for different types of trip destinations. Some of the observed statistically significant results seem to be intuitive. More importantly, it seems that results can reflect some travel preferences that have never been captured in the existing liter-

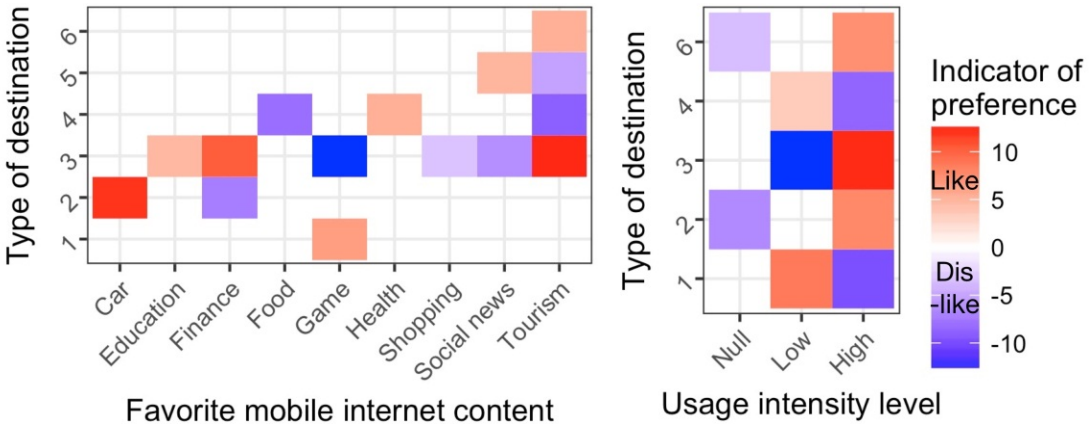


Figure 3.7: The statistical relationships between the mobile internet usage behaviour and the preferences for the types of trip destinations in the initial loop using the original spatial trace.

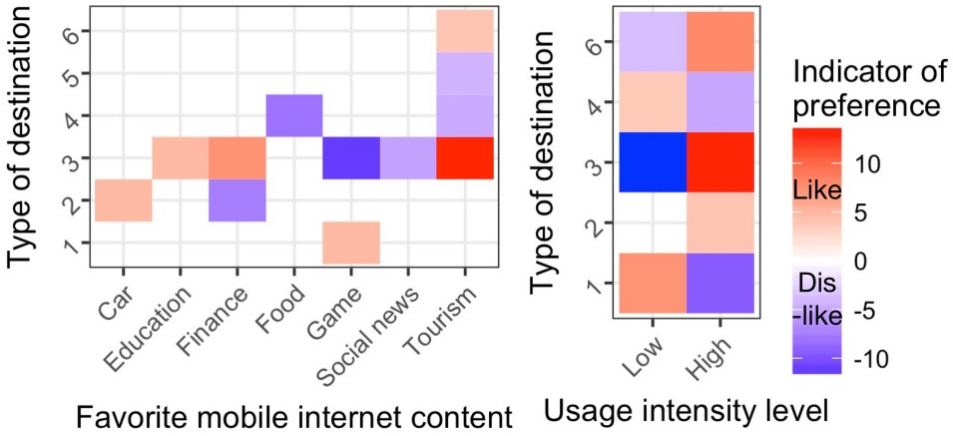


Figure 3.8: The robust statistical relationships between the mobile internet usage behaviour and the preferences for the types of trip destinations in the 20 loops.

ature using traditional travel survey data, mainly because travellers are now grouped based on their specific interests. For example, the destinations of type 6, which are mostly related to touristy areas, were only significantly visited by people tagged with the tourism label. Some existing studies have explored the travel preferences of car lovers, and they found that car lovers had their own preferences for residential or job location choice since it is flexible for them to travel farther, and they are not much willing to drive to downtown often (Van Wee, 2009; Van Wee et al., 2002). In our study, we found that car lovers also had their own preference for secondary activities, which shows that they would significantly like/need to visit the commercial or industrial areas far from the city centre. Despite being more multifunctional, the characteristics of the destination type 3, compared to the other types, are more similar with the concept of a CBD (central business district) since a CBD is usually located in the city centre, also being the most attractive part of the city. It was found using traditional travel survey data that younger people preferred to visit CBDs (Sivakumar & Bhat, 2007). In our study, we assume that users who preferred education and game contents are more

likely to be younger, and they seem to have totally different preferences for the destination type 3. This implies that age might not be sufficient to segment the population for travel analysis, and mobile internet behaviour might be able to reflect a persons characteristics in more detail.

Users with different levels of total mobile internet usage intensity also had clearly different preferences for different types of destinations. Users who often used mobile internet preferred to visit the destinations of type 2, 3 and 6, and they did not visit often the destinations of type 1 and 4. In contrast, users who less often used mobile internet preferred to visit the destinations of type 1 and 4, and they did not visit often the destinations of type 3 and 6. It can be observed in Table 3.1 that cluster 2, 3 and 6 are distinct from cluster 1 and 4, as the former are more commercial, and the latter are only residential. Therefore, we can draw the conclusion that those who used mobile internet intensively were more likely to visit commercial areas, and those who used less mobile internet preferred to make trips to more residential areas. It is worth comparing our results with the results of an existing study using travel survey data (Giuliano et al., 2003). In that study, researchers found no significant difference between the land use of the places where the elderly and the non-elderly travelled. We have similar results in our case, if we assume people who never used mobile internet services were most likely to be the elderly. However, among those who use mobile internet services, the level of usage can be related to their preferred destination types.

It is also worth comparing our results with the relevant ones found in the recent studies using new big data sources. A research group from Estonia was able to access mobile phone traces associated to users demographics, and by using such data, they conducted several studies to investigate the impact of ethnicity, age, and gender on activity locations and spaces (Järv et al., 2014; Silm & Ahas, 2014; Silm et al., 2018). They found in their case studies that ethnicity had a significant influence on the spatial preferences of individuals for out-of-home non-work activities, and the ethnic segregation in activity spaces was higher in younger age groups. We believe that our results can complement the results of those studies since mobile internet usage data (the apps and webpages used) can characterize users in a different way. Also, it is arguably easier to re-identify users by using mobile phone traces associated to users demographics, which is not desirable from a privacy perspective. Our approach seems to be able to distinguish different population segments at a relatively lower privacy risk. As another promising new mobility data source, social media data, including user-generated text, hash-tags, check-in information and even photos, can provide rich contexts of locations and users, allowing researchers to estimate more accurate activity purposes and find more specific interests of users (Hasan & Ukkusuri, 2014; Huang et al., 2017; Rashidi et al., 2017). Thus, it is also possible to characterize users and relate them to mobility behaviour using social media data. For example, Hasan & Ukkusuri (2015) used the Foursquare check-ins posted via Twitter to understand peoples different attitudes and interests through activity locations. Also by using Twitter data, Abbasi et al. (2015) were able to identify the tourists in Sydney and at the same time find that they

were more likely to visit touristy areas. However, an issue of using social media data for such analysis is that the users of a social media product may not be an unbiased sample of the general population of travellers, both demographically and geographically (Hasan & Ukkusuri, 2015), when compared to the general mobile phone users.

3.5 Conclusions and recommendations

This paper proposes a method to segment the population and understand travellers preferences for types of trip destinations by fusing mobile internet usage data and mobile phone traces. The results of a case study, using a dataset from Shanghai, China, show that given ones favourite category of mobile internet content, the proportions of visiting some types of destinations were significantly higher, and the proportions of visiting some others were significantly lower. Many of these observed relationships were interpretable. For example, compared to the others, the users whose favourite content was tourism preferred to visit the touristy areas. Moreover, the users intensively using mobile internet were more likely to visit more commercial areas, and the users who used mobile internet less often would prefer to visit predominantly residential areas.

There are some limitations in this study which derive essentially from the data quality. The sampling rate of the mobile phone traces is relatively lower. As we have discussed in Section 3.3.1, we have to stipulate that at least two consecutive traces close in space can determine a stay point. This will necessarily lead to overlooking some short activities; however, it is the best approach for the available data. In addition, we only use the number of POIs to reflect the characteristics of an area; however, some additional information about the POIs can be added to improve our model. For example, in our case, we found that the number of restaurants is not very different in different areas of the region, but as we know, the quality of restaurants can be very diverse, and it is possible that the better restaurants are spatially distributed in a different way than the others. Thus, although our current model cannot distinguish the areas frequented by foodies, it may be possible to do so, by using more detailed data such as ratings or more specific categories of POIs.

People may also question about the potential privacy issues of such analysis since users generally do not want their mobility traces or mobile usage to be disclosed (Blondel et al., 2015). Despite such privacy risks, society can however benefit from using such big data for transport analysis and planning. Therefore, it is important to consider the extent to which such data should be pre-processed before being available for researchers or decision makers. For example, the data should be aggregated to prevent the privacy risks, whilst at the same time it should not be overly aggregated since it could cause the loss of information and make transport analysis not accurate. In our case, we think that the data provider found a good balance of data aggregation. They aggregated the mobile phone traces hour by hour. They also aggregated the specific websites and apps into several categories, and they only provided the frequency of

each user visiting each category of websites and apps during a period. Even though the demographics of the users were removed, the aggregate mobile internet usage data can still help distinguish different population segments.

The significant and interpretable relationships found in this case suggest the potential of using mobile internet usage data to enhance the explanatory travel behaviour models in future research. Although we only explored the statistical significance in this case, several applications can be made based on the findings of our study. For example, mobile internet usage data may be used to predict mobile users destination choice or for developing a travel behaviour model that would benefit from population partition, such as trip generation model and mode choice model.

3.6 Acknowledgment

We would like to express our gratitude to the Shanghai Unicom WO+ Open Data Application Contest for making the mobile phone data available for this research.

Chapter 4

Nearest-neighbor collaborative filtering for modeling location choice

---

The previous two chapters used proxy variables for personal attributes, either socio-geographic status or mobile internet usage, to segment a given population and further predict spatial behavior. However, one might ask whether it is necessary to make prior assumptions that relate spatial behavior to certain personal attributes. Alternatively, a data-driven approach can be taken under a more flexible assumption: past behavior itself can reflect the heterogeneity in the population and be further used as a reference to predict future behavior. This chapter introduces an algorithm called collaborative filtering, which is likely unfamiliar to many transportation researchers but has been widely used in product recommendation systems. A neighborhood-based collaborative filtering algorithm is tailored to predict non-work location choice. The method is applied to the metro smart card data from Shanghai.

The chapter is based on the following paper that is currently under review:

Wang, Y., Correia, G.H.A., van Arem, B., & Timmermans, H.J.P. (2020). Exploring a neighborhood-based collaborative filtering approach to modeling location preferences for flexible activities through metro smart card data. Journal of Transport Geography, submitted.

Abstract

Given a limited length of time of observation, the frequencies of individuals visiting different locations revealed in big urban mobility data are still too sparse to represent their real location preferences, especially for those non-work and out-of-home activities, which are flexible in space and time and infrequent by nature. Traditional discrete choice models address this issue by imposing theory-based prior assumptions, i.e., the contribution of certain factors to utilities resulting in a location choice. This method has been very well coupled with travel survey data, which contains personal information to distinguish travel behavior of different population segments. While discrete choice models could also be applied to big urban mobility data, this paper proposes a neighborhood-based collaborative filtering approach to understand travelers location preferences for flexible activities. This data-driven method has been used to model the preferences of a consumer for a product in a recommendation system. It only relies on empirical observations, captures location preferences in a flexible way, and intrinsically accounts for behavioral heterogeneity without prior knowledge of any personal attributes, thus being able to prevent privacy issues. The tailored neighborhood-based collaborative filtering algorithm is applied to the metro smart card data from Shanghai, China. We especially focus on those zero observations, i.e., the metro stations that an individual commuter has not visited during the observation period for flexible activities, and we attempt to use our algorithm to predict which of them are most likely to be visited during the test period. Results show that the collaborative filtering algorithm performs reasonably well in this task, giving support for exploring this method further, which is still relatively unfamiliar to most transportation researchers.

Keywords: Location choice; individual mobility; smart card data; collaborative filtering; nearest-neighbor model.

4.1 Introduction

Location choice for an activity is an important dimension of urban travel behavior. People have significantly different preferences in not only the relatively more long-term location choices, i.e., choices of home and work locations (Sermons & Koppelman, 2001; Timmermans et al., 1992; Willigers & van Wee, 2011), but also location choices for non-work and out-of-home activities which are relatively more flexible in space and time (named flexible activities in this study), such as recreation and eating out (Horni et al., 2009). Traditionally, transportation researchers can only observe a very small sample of travelers through mobility surveys (Chen et al., 2016). Using these data, models can further be developed to explain observed location choices based on some features of locations, such as attractiveness and geographic position, as well as some features of travelers, such as socioeconomic attributes (Arentze & Timmermans, 2007). Location choice models can generally be used for two purposes:

- Purpose 1: estimating current location preferences of a given population, without any changes that would result in a different equilibrium,
- Purpose 2: predicting location preferences in a future scenario with a different equilibrium caused by, for example, a policy or infrastructure change.

In recent years, mobility survey data have been criticized for being outdated, small in size and expensive to collect (Demissie et al., 2015; Wang et al., 2019). On the other hand, big urban mobility data, such as mobile phone data, smart card data, and social media data, have received more and more attention in the transportation field. Because of the granularity and size of such data, they seem promising to reveal individual urban travel behavior as well as a whole picture of urban mobility patterns in a cost-efficient and real-time way (Cats et al., 2015; Demissie et al., 2013b, 2016; Gong et al., 2018; Luo et al., 2019; Tao et al., 2014; Tu et al., 2018; Wang et al., 2018).

Today it is feasible to use big data to observe aggregate visiting frequencies or even detailed spatial-temporal traces of everyone in a given population during a certain period; however, this does not mean that location choice modeling has become irrelevant. To achieve Purpose 2, it is apparently still necessary to understand the factors influencing location choice by fitting traditional discrete choice models to big data (see Wang et al. 2017 for example). Regarding Purpose 1, while it seems sufficient to use historical observations about a given population in big data as a proxy for estimating their current location preferences, this is not necessarily true because such observations are sometimes too sparse to represent the real location preferences of individuals, especially given a limited length of time of observation at the current equilibrium.

Evidence is often abundant to validate an individuals choices of home and work locations using big data (Ma et al., 2017; Zhou et al., 2014), and this kind of choices is mostly fixed at least in the short term. On the other hand, for most flexible activities, due to their infrequency in nature and flexibility in both space and time, it is difficult to observe the full spectrum of location preferences, unless the data is available over very long periods, which would, however, be very likely to involve equilibrium shifts. Given an individual who has never traveled to a place in the past months, it does not necessarily mean that he/she has zero preference for this place. In other words, real location preferences for flexible activities are actually distributed more smoothly than observed visiting frequencies. Therefore, estimating real location preferences can help better understand the current equilibrium and predict future choices of locations to perform an activity, especially those that have not been visited during the observation period, as long as the equilibrium remains unchanged.

To predict an individuals future visiting patterns at a stable equilibrium, using historical visiting frequencies of this individual would cause the problem of overfitting, while in contrast, using historical visiting frequencies of the whole population would cause the problem of underfitting. It is necessary to find a balance between underfitting and overfitting (Calabrese et al., 2013). Discrete choice models tackle this balance by

making behavioral assumptions. One of them is the assumption that people of different socioeconomic strata have different behavior.

For privacy reasons, big urban mobility data are usually anonymous, which means that personal attributes of individuals are always missing in such data, leading to the lack of sources to understand behavioral heterogeneity (Calabrese et al., 2014). (Wang et al., 2017, 2018) attempted to bridge the gap by using proxy variables, including ones socio-geographic status and mobile internet usage, to segment the population and further explain their different travel behavior expressed through mobile phone traces and smart card data.

Alternatively, a data-driven approach can be implemented under a much more flexible assumption (He et al., 2015): past behavior itself can reflect the heterogeneity in the population and be further used as a reference to predict future behavior at a stable equilibrium. This approach is very similar to a computer-science concept that has been widely used in product (e.g., movie or music) recommendation systems but has rarely been mentioned in the transportation field: collaborative filtering (Schafer et al., 2007). This suggests that the preferences of an individual can be modelled by collecting preference information from many other similar individuals.

An analogy can be made here between modeling movie preferences in a recommendation system and modeling travel location preferences. In the past, computer scientists created explicit profiles to characterize users (e.g., age) and movies (e.g., genre) and then built statistical models to explain their relationships and further make predictions. For example, young people would prefer thriller movies. This approach was termed as content filtering (Su & Khoshgoftaar, 2009). However, such preferences are complicated in most cases, and many influencing characteristics of users and movies are actually too latent and subtle to observe. Also, it is inefficient to collect so much explicit information about so many users and movies.

With the emergence of big data and the expansion of computational capabilities, computer scientists started taking the approach of collaborative filtering, which does not require knowing any explicit information about users and movies and only relies on historical interactions between users and movies (Koren et al., 2009). In fact, the traditional approach of location choice modeling is pretty much the same as content filtering, where travelers and locations are the equivalents of users and movies. Transport modelers collected explicit information about travelers and locations, and then tried to explain travelers location preferences (Arentze & Timmermans, 2007). For example, young people would prefer crowded areas.

We argue that collaborative filtering is promising for modeling location preferences mainly for two reasons. First, big urban mobility data and high computational capability are becoming accessible to the transportation research community as well (Chen et al., 2016). Second, it is inefficient and even impossible (for privacy concerns) to collect explicit information about travelers and locations in big data. Also, many characteristics of travelers and locations that influence location preferences are usually

latent, subtle and thus difficult to explicitly incorporate in a discrete choice model. Instead, collaborative filtering approaches can benefit from massive historical data and intrinsically account for behavioral heterogeneity.

In general, collaborative filtering can be implemented as either a nearest neighbor model or a latent factor model (Koren et al., 2009). We focus on the former in this study and leave the latter for future research. Collaborative filtering with a nearest-neighbor model is also called as neighborhood-based collaborative filtering, and the underlying logic of this method for our problem is straightforward: the full spectrum of ones location preferences can be approximated by considering the behavior of ones neighbors, and here the neighbors are defined by the similarity between their historical behavior.

This concept is tested in the case study of the metro smart card data from the city of Shanghai, China. We especially focus on those zero observations, i.e., the metro stations that an individual commuter has not visited during the observation period for flexible activities, and we attempt to use our algorithm to predict which of them are most likely to be visited during the test period. As explained, this topic might not be the most interesting one to the people whose objective is mainly to estimate the elasticity of travel demand with respect to operation, infrastructure or policy changes (Gan et al., 2020; Lin et al., 2018). However, it would be relevant for those who want to understand individual mobility patterns at the current transport demand-supply equilibrium (Axhausen et al., 2002; Borgers et al., 1989; Sivakumar & Bhat, 2007; Yin et al., 2017; Zhao et al., 2018). Also, predicting such location choices for flexible activities is more challenging than predicting regular commuting trips or self-repetitive non-commuting trips (Goulet-Langlois et al., 2017) and has thus attracted many researchers attention (Danalet et al., 2016; Horni, 2013; Marchal & Nagel, 2005). To the authors knowledge, it is the first time that this topic has been studied following a collaborative filtering approach.

The remainder of this paper is organized as follows. First, the case study of Shanghai is described, with an explanation of the data preprocessing. Then, the methods are proposed, with some benchmarks in order to give readers a term of comparison, such as a simple discrete choice model. Following that, the results are presented and discussed. Finally, conclusions are made, and future research directions are pointed out.

4.2 Case study and data preprocessing

The smart card data provided by the Shanghai Open Data Applications (SODA) contest[1] contain the records of all transactions by all smart cards in Shanghai, China from July to September, 2016. Metro is the only public transport (PT) system in Shanghai where cardholders should both check in and check out. In contrast, travelers must scan

[1] http://soda.shdataic.org.cn/

their cards only when boarding a bus or alighting a taxi, and the location information is missing on these two modes. Therefore, for further analysis, this work focuses on the transit travelers who have only used the metro during the study period. Nevertheless, the methodology can be generalized for other PT systems as well if there are available data.

In addition, only a few metro stations require travelers to check out and then check in again to switch to another line in Shanghai. Such cases should not be regarded as two separate trips. To distinguish them, a threshold of 30 min is used between check-out and check-in at those stations. The selection of this threshold is based on the policy by which after 30 min without checking in again, the system will regard the next check-in as the start of a new trip and will charge for a new trip. Travelers are assumed to be aware of this fact, and if they stay at those stations for more than 30 min, they must have performed an activity whose utility can compensate for the added cost.

Since this study focuses on modeling location preferences for flexible activities, we need to detect a metro travelers home and work stations, if any, by using a rule-based algorithm. The details of this algorithm can be found in the work by Wang, de Almeida Correia, de Romph, & Timmermans (2017). It is not specifically explained here since it is not the main focus of the present work. A metro traveler with detected home and work stations can be defined as a metro commuter. It is assumed that metro commuters perform activities between trips. The following steps are taken to tentatively detect activity purposes of trips within one day:

- If the check-out station of one trip and the check-in station of the next trip are the same one, the activity purpose can be classified into home, work or flexible activity, depending on whether the station is the home, workplace, or neither for that individual.
- The check-out station of one trip and the check-in station of the next trip can also be different due to the intermediate unobserved movement using other modes of transport. In that case, the activity purpose is labeled as undefined.
- The first activity in a day is dependent only on the check-in station of the first trip, and the last activity is dependent only on the check-out station of the last trip.

As a result, we can extract the number of trips by each metro commuter visiting each station for flexible activities. The specific task of this study is to predict for each metro commuter the station choices for flexible activities during the test period, which have not been visited during the observation period, given their station choices for flexible activities during the observation period. We set July and August, 2016 as the observation period and set September, 2016 as the test period. Figure 4.1 shows an example where the individuals home and work stations are indicated by a star and a triangle respectively, and the number of visits to the other stations for flexible activities in the three months is indicated by the squares with different depths of color.

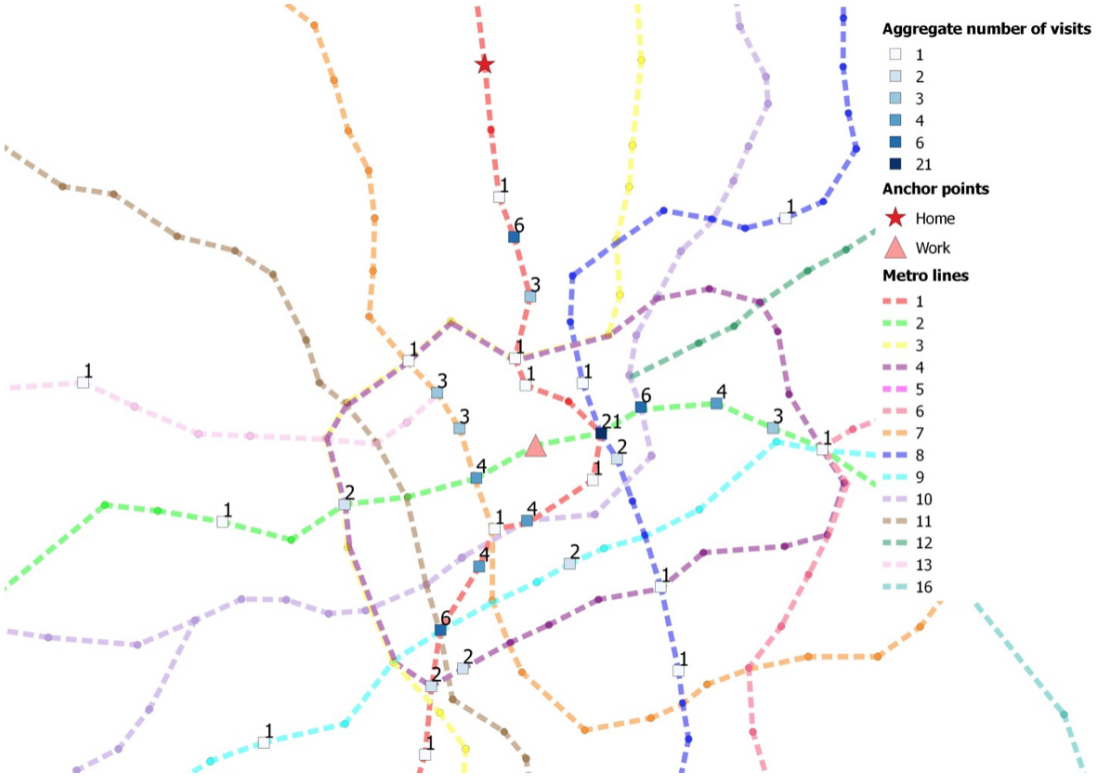


Figure 4.1: The spatial distribution of a Shanghai metro commuter's flexible activities in three months.

4.3 Methods

The approach of collaborative filtering does not require making any theory-based prior assumptions, and it only finds behavior patterns in empirical data. The model mainly consists of two steps (Aggarwal, 2016). The first step is looking for the neighbors which are individuals who share similar historical behavior patterns to a specific traveler i. In this work, the data during the observation period are used to determine neighbors and generate predictions. For a traveler i, a vector $\boldsymbol{f}_i$ is constructed to represent the frequency $f(i,s)$ of traveler i visiting each station s during the observation period. The similarity between two travelers historical behavior patterns is defined as the Pearsons correlation between the frequency vectors of them.

An important issue is to define the range of what is considered to be a neighbor. A common approach is to select a certain number of neighbors, which is known as a k-nearest neighbor model (Aggarwal, 2016). Instead, in our study, we select the neighbors by applying a cutoff value of the correlation-based similarity because in this way, it is easier to interpret the similarity between the neighbors and a traveler based on whether the correlation coefficients are positive or negative.

The range of the cutoff value is set to vary between 0 and 0.5 at a 0.1 interval. Zero correlation is a reasonable and conservative cutoff value just by definition. In addition,

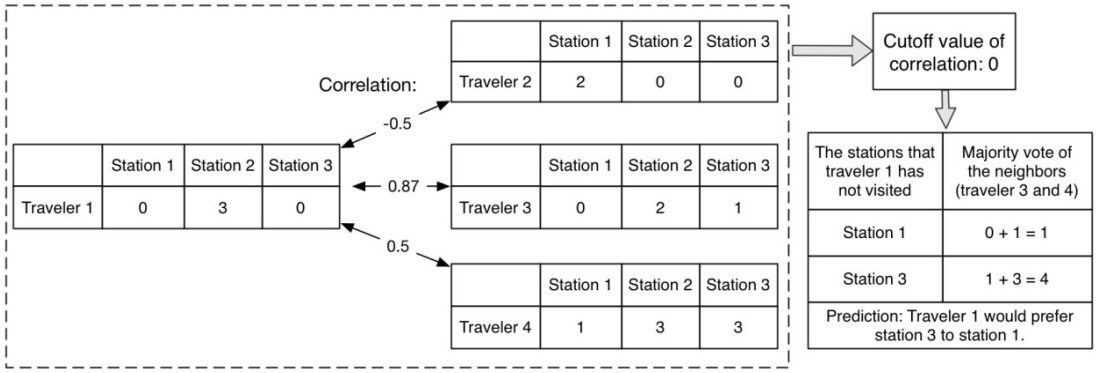


Figure 4.2: A toy example of the neighborhood-based collaborative filtering algorithm for flexible activity location choice prediction.

if the cutoff value is set as -1, the collaborative filtering model will actually collapse to the method in which it is assumed that people always choose the station most visited by all other travelers. In this paper, it will be tested how different cutoff values influence the prediction accuracy. This issue is equivalent to finding the optimal k (i.e., considered number of neighbors), regarded as a hyper-parameter, in a k-nearest neighbor model (Hastie et al., 2009).

The second step uses the majority vote of the neighbors of a traveler to calculate a prediction for that traveler. This study aims to predict the station choice for flexible activities that a traveler has not visited during the observation period; therefore, the vote refers to the total frequency of visiting each station that traveler i has not visited, by the neighbors of traveler i in the past. A vector of neighbors vote regarding traveler i can be constructed as $\boldsymbol{m}_i$, where each element m_i is equal to $\sum_{N_i} f(n_i, s_i)$, and s_i is a station that traveler i has not visited in the first two months; a neighbor of traveler i is expressed by $n_i \in N_i$. Next, the top ranking stations are chosen for traveler i, which are the ones that the neighbors of i have visited most times. These are regarded as predictions of the choice by traveler i.

A toy example of the model is shown in Figure 4.2. It is assumed that there are four travelers in a network of three stations. The correlation-based similarities of traveler 1 to the other travelers are calculated using the historical records, and traveler 3 and 4 turn out to be the neighbors of traveler 1. Based on the majority vote, it is predicted that traveler 1 would prefer station 3 to station 1.

The data of the test period are used for evaluating the predictions. If the predicted station matches one of the actual stations that traveler i has visited during the test period, a hit is achieved.

In this study, we use a simple multinomial logit discrete choice model as a benchmark in order to give readers a term of comparison for how the collaborative-filtering approach performs. We want to remind again that our research objective at this stage is only to explore the possibility of the new approach. Discrete choice models are still irreplaceable.

Traditionally, discrete choice models have been used to model location choice for flexible activities considering the potential utilities of choosing available locations, which are dependent on certain factors (e.g., attractiveness, distance, and personal attributes). Given the available metro smart card data in our case study, the deterministic part of the utility function of choosing a metro station, among all the stations in a metro network, by a metro commuter for a flexible activity can simply be expressed by the following equation:

$$v(i,s) = \alpha A(s) + \beta_1 hd(i,s) + \beta_2 wd(i,s) \tag{4.1}$$

Where $v(i,s)$ is the deterministic utility of choosing station s for a flexible activity by traveler i, where s should neither be home station nor work station for i; $A(s)$ is the attractiveness (usually in terms of size) of station s, and more specifically, the number of POIs within the radius of 500 m around station s. $hd(i,s)$ is the metro network distance between the home station and station s, and $wd(i,s)$ is the metro network distance between the work station and station s (in terms of km). α is the parameter representing the preference for attractiveness; β_1 and β_2 are the parameters representing the preference for distance from home and distance from work respectively. Personal attributes are not available for this model as they are missing in smart card data. Therefore, it is difficult to distinguish the preferences between different population segments. In addition, it was found in our previous work (Wang et al., 2017) that the effect of spatial autocorrelation is insignificant in this kind of choice situations; thus, it is not incorporated in this model.

Despite being simple, the model is already able to capture the most basic factors of location choices: size and distance of a location. For each traveler, we use the estimated parameters to calculate the deterministic utility of choosing each station, and based on that, we rank the stations and exclude those that have been visited by this traveler during the observation period. Among the remaining stations, the one with maximum utility is regarded to be the prediction generated for this traveler.

Several other benchmarks are used to generate top ranking stations as well, which are compared with the ones predicted by the nearest neighbor method proposed in this paper:

- Assuming that people always choose the station that is most visited by all other travelers (equivalent to a collaborative filtering algorithm with the cutoff value of -1);
- Assuming that people always choose the station with most points of interest (POI);
- Assuming that people always choose the station closest to home;
- Assuming that people always choose the station closest to work.

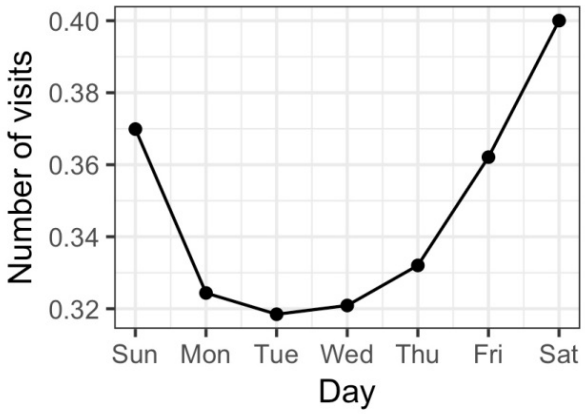


Figure 4.3: The average number of visits per metro commuter for flexible activities on each day of week.

4.4 Results

In the case study, we focus on the metro commuters who left intensive travel records (i.e., traveled at least two-thirds of the days) and made a trip for flexible activities at least once in the first two months and at least once in the third month. On average, these travelers visited around 5 unique stations for flexible activities in the first two months and visited around 3 unique stations for performing flexible activities in the third month. Figure 4.3 shows the average number of visits per metro commuter for flexible activities on each day of the week.

In the third month, about 86% of the preselected travelers visited at least a new station that has not been visited in the first two months. This indicates that our hypothesis is somehow true in this case: the full spectrum of location preferences cannot be fully observed in the first two months. We further focus on these 37,923 travelers for the prediction task.

For each traveler, we look for the neighbors among the 37,922 travelers (excluding the analyzed traveler) based on the similarity in terms of the correlation coefficient, with a cutoff value of the correlation coefficient (as explained in Section 4.3). If a neighbor is specifically defined as the person whose similarity is over 0, each traveler would have 25% of all other travelers as his/her neighbors on average according to the data. With the increase of the cutoff value, the percentage of neighbors decreases, and it decreases to an average of 1%, if the cutoff value is set as 0.5.

For performance benchmarking, using the data of the first two months, we estimate a discrete choice model whose deterministic utility function is given by Equation 4.1. The estimation results are presented in Table 4.1. It is not a surprise to observe that travelers generally prefer a closer station with more POIs. In addition, on average, they seem to prefer one closer to home than from work.

Figure 4.4 shows the number of hits of all the considered methods tested on 37,923

Table 4.1: The estimation results of the multinomial logit model.

Parameter	**Coef.**	**Robust Std. err.**	**Robust t-test**
α (preference for number of POIs)	4.71e-04	3.93e-06	120.04
β_1 (preference for distance from home)	-0.132	0.00102	-129.39
β_2 (preference for distance from work)	-0.0134	0.00121	-11.06
Number of observations: 37,923; Adjusted rho-square: 0.129			

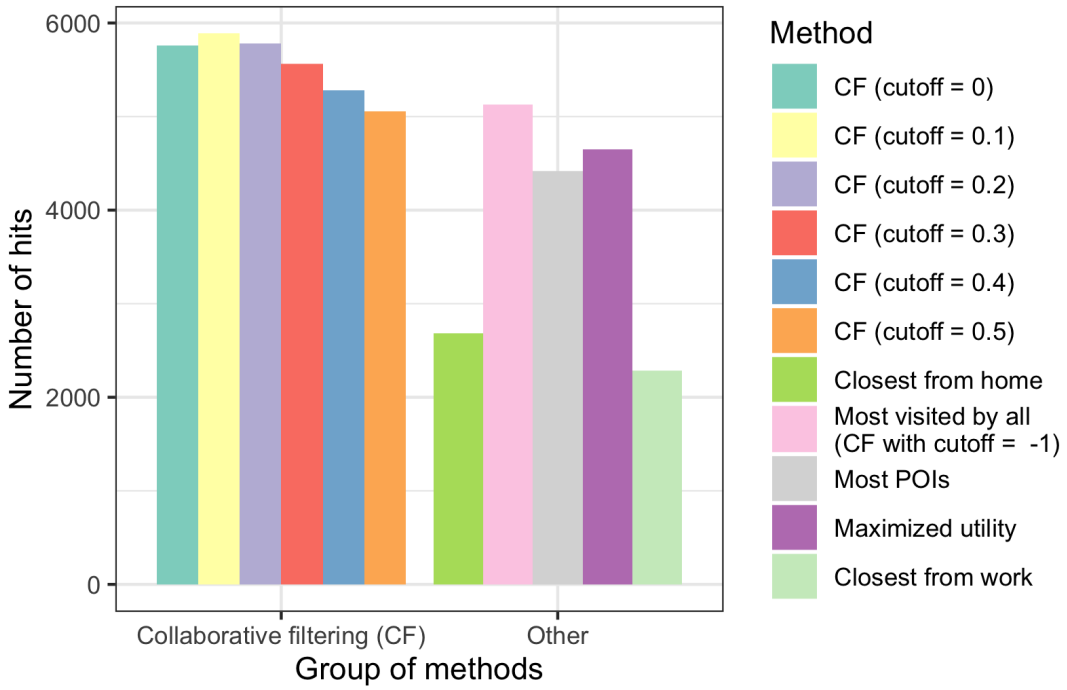


Figure 4.4: The prediction results of all the methods applied to the 37,923 travelers.

travelers. A hit is defined as the situation where the predicted station matches one of the actual stations to be visited in the next month. First of all, and most importantly, the collaboration filtering method with almost all the cutoff correlation thresholds outperforms all the non-collaborative-filtering methods. The best collaborative filtering method results in nearly 15% more hits than the method assuming that people choose the station most visited by all other travelers, and about 16.5% more than the estimated multinomial logit model. Consequently, the number of correctly predicted choices would further increase in a scaled-up scenario using our collaborative filtering methods. The hit rates (about 6,000 hits among 37,923 travelers) do not seem very high overall; this is actually not surprising because the specific problem, predicting location choice of flexible activities, which have not been visited before, is not an easy task and must be more challenging than predicting other more regular location choices.

Among the non-collaborative-filtering methods, the hit rate of the utility-maximization-

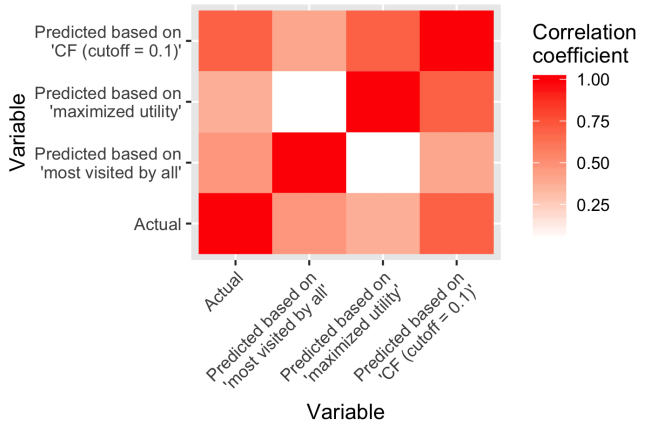


Figure 4.5: The correlations between the actual number of visits to a station in the third month vs. the number of travellers who are predicted to prefer the most this station by different methods.

based multinomial logit model is slightly lower than the hit rate of the method assuming that an individual traveler always chooses the station that is most visited by all other travelers. Moreover, the number of POIs seems to be more influential than the other two single features used in the multinomial logit model (i.e., distance from home and distance from work).

While the number of hits reflects the individual-level accuracies of different methods, to check the station-level accuracies, we calculate the correlations between the actual number of visits to a station and the number of travelers who are predicted to prefer the most that particular station, as shown in Figure 4.5. It can be observed that among the other methods, the collaborative filtering method with a cutoff value of 0.1 can generate a prediction that has the highest correlation (around 0.7) to the actual number of visits.

Furthermore, we visualize the actual vs. predicted location preferences among the top-50 most visited stations on the map of Shanghai. Here, location preference is defined as a normalized indicator that represents peoples relative preference to visit each station. For the actual observations, it is the number of visits to a station divided by the total number of visits. For the predictions, it is the number of travelers who are predicted to prefer one station the most divided by the total number of travelers. As shown in Figure 4.6, the predictions made by the most visited by all method are the least smooth ones. Except for a few stations that are predicted to be very popular, the preference for the remaining stations is mostly under-predicted. While the predictions made by the collaborative filtering method and by the discrete choice model assuming utility maximization are both smooth, it seems that the former is more accurate (e.g., the station at the most left side).

The results imply that the collaborative filtering method might be able to account for some latent and subtle effects that the other methods cannot capture. To gain such insights from the prediction results for a better interpretation, we explore the extent to

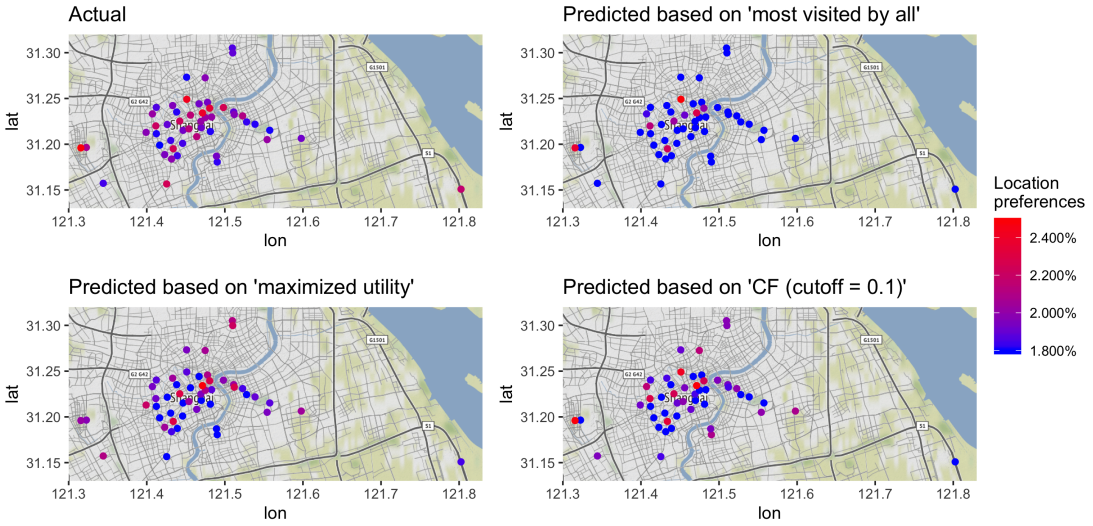


Figure 4.6: Actual location preferences vs. predicted location preferences among the top-50 most-visited stations.

which a metro station s visited in the first two months would, in part, lead to the correct prediction of the station s' to be visited in the third month, by counting the transition frequency from s to s' in all the correctly predicted cases using collaborative filtering with the cutoff value of 0.1.

The transition with the highest frequency is from Shanghai Railway Station to Hongqiao Railway Station. Note that they are both the names of the metro stations connected to the corresponding railway stations. This means that according to the collaborative filtering algorithm, people are likely to visit Hongqiao Railway Station in the third month, given that they have visited Shanghai Railway Station in the first two months. These two metro stations are distant from each other, but they obviously have the same function. It seems that our collaborative filtering method can intrinsically capture such preferences shared by the frequent inter-city travelers, possibly business people. To show how the results vary with different random samples and further validate the optimal cutoff value, we randomly sample three groups of 500 users. In each group, the prediction results of the collaborative filtering methods as well as the most visited by all method (equivalent to a collaborative filtering algorithm with the cutoff value of -1) are shown in Figure 4.7.

It can be observed that in each sample group, with the increase of the sample size, the line describing the relationship between hit rate and cutoff value always converges to a similar pattern: hit rate achieves the highest as cutoff value ranges from 0 to 0.2. The same pattern was also found in Figure 4.4. Accordingly, it can be concluded that the optimal cutoff value should neither be too high nor too low. It is relatively understandable that the cutoff value should not be too low because otherwise one would have an increasing number of neighbors who are actually not similar among themselves. On the other hand, it is at first glance counterintuitive to find that 0.5 is not the best cutoff value in our case, but it actually makes sense because it has been found that a too small

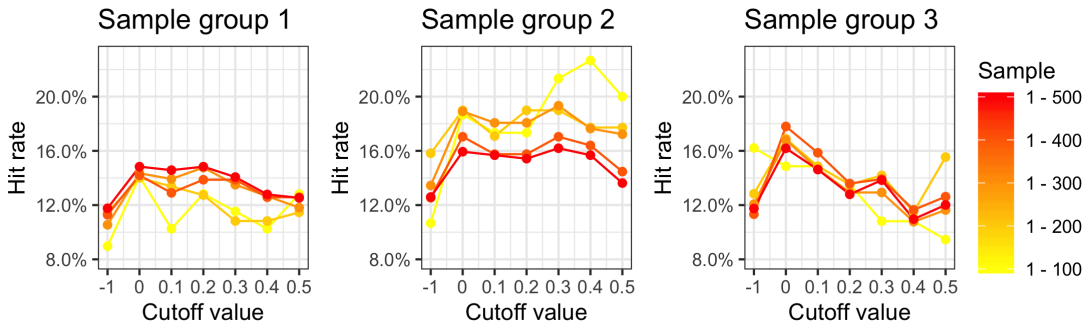


Figure 4.7: The prediction results of the collaboration filtering methods applied to different groups of travelers.

k would potentially lead to unreliable predictions due to overfitting (Altman, 1992).

4.5 Conclusions and recommendations

This paper proposes a data-driven method, namely a neighborhood-based collaborative filtering algorithm, to model location preferences of individuals for flexible activities (i.e., those non-work and out-of-home activities), which are too sparse to observe even in big data. Different from discrete choice models, which can smoothen observed discrete choices by imposing theory-based prior assumptions regarding travel behavior, our approach can smoothen observed discrete choices using a nearest-neighbor model, which approximates behavior patterns merely based on empirical evidence, being thus very suitable for big mobility datasets. One of the advantages of such smoothing is the ability to estimate the full spectrum of location preferences. Even though a location has not been visited by an individual during the observation period, it does not necessarily mean that the individual has zero preference for this location. Therefore, for each individual, we specifically focus on those non-visited locations and define our task as predicting which of them are most likely to be visited. The case study is conducted in Shanghai, China, and focuses on the active metro commuters, using three-month smart card data. The results show that the collaborative filtering approach is comparable with the other approaches, in terms of the prediction performance. At least, it shows promise as another possibility of understanding individual mobility patterns at a stable transport demand-supply equilibrium.

Despite being out of the scope of this paper, a few issues are worth discussing and exploring in future research. Our case study focuses on active travelers only, whose intensive historical travel records can help characterize each individual and find the neighbors. However, in many cases, it is also necessary to predict the behavior of the users who have left few travel records and even new users with no records. The same applies to the supply side. The current model cannot predict the preferences of existing users for a newly built station. In the product recommendation field, such problems are designated as cold start problems (Su & Khoshgoftaar, 2009). It is relevant to study

the cold start problems in the context of location choice modeling. It is very likely that we have to make more prior assumptions again to compensate such sparsity of data. Moreover, our research is somehow limited by the data we have. Further in-depth analysis can be done if it is possible to access a mobility dataset over a longer period. We may even consider location choice by different modes if we have multi-modal travel demand data.

4.6 Acknowledgment

We would like to express our gratitude to the Shanghai Open Data Applications (SODA) contest for making the data available for this research.

Chapter 5

Matrix factorization for modeling spatial interactions

Data privacy is a concern if individual spatial traces are tracked. Aggregation is the safest solution. Thus, this chapter assumes that in some cases, big mobility data are only available in an aggregated form, where individual spatial behavior is aggregated into spatial interaction matrices. A matrix factorization approach is proposed to understand the observed spatial interactions. As a continuation of the collaborative-filtering strategy in the previous chapter, this method does not require any prior definition about each location, and learns the production and attraction of a location in a data-driven way. The method is applied to model the spatial interactions between metro stations in Shanghai, China by using an origin-destination trip matrix constructed based on one-day smart card data.

The chapter is based on the following paper that is currently under review:

Wang, Y., Correia, G.H.A., van Arem, B., & Timmermans, H.J.P. (2020). A matrix factorization approach to modeling trip generators and their interactions. Travel Behaviour and Society, submitted.

Abstract

Increasingly available big mobility-related data, such as mobile phone traces and smart card data, show human activities and mobility patterns at a large scale. To explore the application of new methods that are more appropriate for such new big data, our paper attempts to bridge the gap between two techniques from different research areas: (1) unconstrained gravity model, traditionally used for trip distribution, and (2) hierarchical Poisson factorization, a variant of machine learning matrix factorization methods, which has been commonly used to predict user preferences for a product in a recommendation system. We show how a traditional gravity model can be adapted by representing production and attraction in multiple latent dimensions and estimating them in a data-driven way. We also show how a hierarchical Poisson factorization framework can model mobility patterns only by additionally considering the effect of travel costs. With the added extensions, the two methods become equivalent, resulting in a gravity-based Poisson factorization model, which is suitable for modeling urban trip generators and their interactions using big mobility-related data. Using metro smart card data from Shanghai, China, results show that the new model benefits from adding the number of latent dimensions and outperforms the one-dimensional model, decreasing the root-mean-squared error of the test set by up to 34%. It also benefits from the consideration of travel costs, especially with a small number of dimensions. More importantly, it allows predicting mobility flows given new travel cost matrices.

Keywords: Big data; machine learning; matrix factorization; hierarchical poisson factorization; gravity model

5.1 Introduction

Various kinds of big data are available nowadays to capture and analyze human mobility patterns, particularly for urban areas (Hasan et al., 2013a; Noulas et al., 2012). Such data are inherently disaggregate and thus reveal individual movements between locations (Calabrese et al., 2013). These data can be processed and presented in an aggregated form, such as spatial interaction matrices (Zhao et al., 2016a), interchangeably called origin-destination matrices (OD matrices). Although these data help reveal visiting patterns (Hasan et al., 2013b), in many situations, urban planners and mobility-related companies require mobility models with sufficient explanatory and predictive power to determine how different factors contribute to mobility performance (Çolak et al., 2015). For example, a model can quantify the impact of travel distance on spatial interactions. Such models can further be used to support smarter decisions on urban development, transportation network design, and business relocation (Batty, 1976).

To build an explanatory mobility model, it is important to characterize locations, which produce, and at the same time, attract different kinds of travelers for different purposes.

Based on prior knowledge, certain predefined variables (e.g., number of points of interest in a certain category) can be used to differentiate the locations, and the values and/or levels of these variables can be used to explain and/or predict mobility patterns. However, with the increasing resolution of mobility data, it might become more difficult to capture location differences only by predefined variables. A data-driven approach is therefore needed.

Traditionally, transportation modelers developed gravity models and inferred several model parameters, including the effect of generalized travel costs, and trip generators, i.e., production of origins and attraction of destinations, from observed inter-zonal travel flows coupled with their corresponding generalized travel costs (Sen, 1986). Estimated models were then used to predict travel flows in what-if scenarios. In such models, given a single generalized travel cost, travel flows between locations must be higher, as long as the production of the origin and the attraction of the destination, which are represented by some predefined variables such as the population density of a location, are both higher. This is however not necessarily true: two locations can be very densely populated in general but at the same time have little interaction with each other just because of some latent reasons (e.g., hipster vibe) which are not known in advance and only revealed in the data.

Is it possible to account for large-scale spatial interactions in a data-driven way instead? The answer to this question may be inspired by what machine learning (ML) researchers do when they attempt to model the interactions between users and movies (or any other form of products) in a recommendation system, where it is impracticable to collect explicit information about innumerable users and movies. To solve this problem, matrix factorization techniques have been developed to capture multidimensional latent factors of both users and movies, merely based on historical watching or rating patterns. These data-driven models assume that a higher correspondence (mathematically, a higher dot product) between the latent factors defining user-profiles and the ones characterizing a movie would lead to a higher chance of matching (Koren et al., 2009). The same approach could potentially be adopted for mobility-related data, capturing production and attraction without predefining any variable for them, merely based on an observed spatial interaction matrix.

According to this principle, we assume that a high correspondence between a locations multidimensional production factors and another locations multidimensional attraction factors leads to a stronger spatial interaction between these two locations. At the same time, we can borrow the gravity concept by assuming that spatial interactions depend not only on production and attraction factors but also on the impedance of traveling between two locations. Based on this assumption, our study essentially extends a matrix factorization framework with a travel cost function to model the observed spatial interaction matrix of a city. Our method is thus rooted in the domain knowledge of transportation, and it leverages state-of-the-art ML techniques. Compared to a traditional gravity model, the proposed method estimates more latent factors that are not predefined to describe the production or attraction of a location in a data-driven way.

Compared to matrix factorization techniques for recommendation systems, the proposed method additionally accounts for the effect of generalized travel costs.

The approach will be illustrated in this paper using the Shanghai Open Data Applications[1] that provided the check-in and check-out data of Shanghai smart card users at all the metro stations on April 1, 2015. These disaggregate data were aggregated into a station-level daily OD matrix. Our method can be applied to spatial interaction matrices built based on any mobility data sources, including not only smart card data (Zhao et al., 2007), but also cellular network data (Çolak et al., 2015), social media check-in data (Gong et al., 2019), and GPS data (Rasouli & Timmermans, 2014; Wolf et al., 2004). The method presented herein only requires mobility data in aggregate form, thus preventing potential privacy issues (Blondel et al., 2015).

The remainder of this paper is organized as follows. Section 5.2 presents a literature review of traditional gravity models for trip distribution and matrix factorization methods for recommendation systems. Section 5.3 illustrates the method. Following that, Section 5.4 illustrates the method through the use of metro smart card data, and the results are presented and discussed in Section 5.5. Finally, the main conclusions are drawn, and future research directions are pointed out in Section 5.6.

5.2 Background literature

5.2.1 Single-dimensional unconstrained gravity models

Spatial interaction is reflected in the number of trips from one place to another. In the era of big data, spatial interaction has become a more general term, not only for physical human movements but also for some trip proxies that can only be observed in big data, such as two sequential social media check-ins from one location to another (Liu et al., 2014).

From the perspective of a traditional 4-step transportation model (de Dios Ortúzar & Willumsen, 2011), each zone can produce and attract trips. For example, a zone can produce more trips on a weekday morning because it is a residential zone. This is the first step called trip generation. Then, the trips produced from each origin are distributed to their destinations based on the gravitational law, proportional to the trip potential of each zone and inversely proportional to the travel impedance, resulting in an OD matrix. This is the second step called trip distribution.

A regression model is typically used to predict the number of trips produced from a zone and/or attracted to a zone by some factors such as the number of households (Kassoff & Deutschman, 1969). The output of this step serves as constraints to the trip matrix totals by column and/or row in producing a final OD trip matrix. The trip distribution model can either be doubly constrained, if both produced trips and attracted

[1] http://soda.shdataic.org.cn/

trips are known, or singly constrained, if only one of them is known (Fotheringham & O'Kelly, 1989). Otherwise, the first two steps can be expressed through an unconstrained gravity model (Cesario, 1975), which is the focus of this paper. In this case, spatial interaction F_{od} from an origin o to a destination d can be modeled using the following equation:

$$F_{od} = \alpha E_o A_d f(c_{od}) \tag{5.1}$$

where F_{od} and c_{od} (i.e., generalized travel costs) are observed; α is a scaling factor to be estimated (Sen, 1986), and E_o and A_d are production factor and attraction factor respectively. These two trip generation factors can fully act as the parameters to be estimated (Griffith & Fischer, 2013). Then, for each origin (or destination), a production (or attraction) parameter needs to be estimated. The drawback of this approach is that the estimated model cannot be applied to new zones for which production and attraction factors are still unavailable. The two trip generation factors can also be expressed in terms of known trip potential variables. For example, E_o and A_d can be the population at zone o and the population at zone d respectively. Such a model is easy to estimate and interpret, by sacrificing some degrees of freedom in the model. Some other models have E_o and A_d as power functions of known trip potential variables (Fotheringham & O'Kelly, 1989).

In addition, $f(c_{od})$ is a function of generalized travel costs to model the negative impact of traveling between one area and another. The most common functions include the exponential deterrence function $e^{-\beta c_{od}}$, and power deterrence function ${c_{od}}^{-n}$, where β or n needs to be estimated (Hyman, 1969).

5.2.2 Matrix factorization methods

ML researchers in the field of recommendation systems study the interaction between a user and a movie (or any other form of product). Similar to the unconstrained gravity model where an origins production factor is multiplied by a destinations attraction factor to account for the interaction, in a recommendation system, a users preference factor is multiplied by a movies attribute factor. In formal terms, the major difference is that a user us preference factor $\boldsymbol{\delta}_u$ and a movie is attribute factor $\boldsymbol{\mu}_i$ are allowed to be multidimensional, sharing a joint latent space. The interaction between a user and a movie should then be expressed as a dot product, $\boldsymbol{\mu}_i^T \boldsymbol{\delta}_u$. Since the two factors are assumed to be fully latent, they need to be estimated given the observed user-movie interaction matrix $\boldsymbol{R}_{ui}$. In the classical matrix factorization method, the mathematical problem is to find the $\boldsymbol{\mu}_i$ and $\boldsymbol{\delta}_u$ that can minimize $(\boldsymbol{R}_{ui} - \boldsymbol{\mu}_i^T \boldsymbol{\delta}_u)^2$ (sometimes plus a regularization term that can avoid overfitting). Koren et al. (2009) provide a detailed guide to classical matrix factorization techniques.

In this paper, we are interested in a specific variant of matrix factorization methods, hierarchical Poisson factorization (Gopalan et al., 2015). Two key features distinguish this method from other matrix factorization techniques: being Poisson and being hierarchical. First, each cell of the observed matrix is assumed to be drawn from a Poisson distribution and is, therefore, more suitable for expressing implicit interaction frequencies. Second, the hierarchical structure can first model general potential production (or attraction) of a location and based on that, specific multidimensional production (or attraction) factors can be modeled (Gopalan et al., 2015). This fits well with the fact that a location has both a level of general popularity and levels of specific popularity for different demand segments.

Different from the classical matrix factorization method with a deterministic approach, Poisson factorization is a probabilistic Bayesian model (Gelman et al., 2013). The model essentially aims to infer the distribution $P(\boldsymbol{\mu}_i, \boldsymbol{\delta}_u | \boldsymbol{R}_{ui})$, the posterior probability of the latent variables, given the available data, based on Bayes theorem as shown in the following equation:

$$P(\boldsymbol{\mu}_i, \boldsymbol{\delta}_u | \boldsymbol{R}_{ui}) = P(\boldsymbol{R}_{ui} | \boldsymbol{\mu}_i, \boldsymbol{\delta}_u) P(\boldsymbol{\mu}_i, \boldsymbol{\delta}_u) / P(\boldsymbol{R}_{ui}) \tag{5.2}$$

Inference is commonly based on two mainstream methods: Monte Carlo Markov chain and variational inference (Gelman et al., 2013), which both require extensive computational power, since they require the estimation of the full posterior distribution. Otherwise, a simpler approach can be taken: Maximum a posteriori (MAP), which estimates the values of $\boldsymbol{\mu}_i$ and $\boldsymbol{\delta}_u$ that maximize the posterior distribution. This estimation method obtains a point estimate, and hence does not require the estimation of the full posterior distribution.

5.3 Methodology

First, we articulate the assumptions and the problem of this study as follows. We assume that the given observations include an OD matrix $\boldsymbol{F}$ and a matrix of generalized travel costs $\boldsymbol{C}$. Moreover, we assume that the OD matrix is dependent on (i) a function of generalized travel costs $f(C)$, (ii) general and specific production factors γ and ρ respectively, and (iii) general and specific attraction factors φ and ω respectively. The goal is to estimate these factors and the parameters of the assumed generalized cost function given the observed data. Once the model has been estimated, it can be applied to what-if scenarios with a new matrix of generalized travel costs as the model input, and the potential spatial interactions as the model output.

The spatial interaction matrix to be factorized $\boldsymbol{F}$ is an $S \times S$ square matrix. There can be observations on the diagonal if self-interaction is considered. It can otherwise be a matrix without any observation on the diagonal if self-interaction is not considered. A

Gamma-Gamma hierarchical structure is used to model the production and attraction factors because it has been found to fit well the possible skewness in data and contribute to better predictive performance (Gopalan et al., 2015). At the top of the hierarchy, general potential production, and general potential attraction are modelled first for each location. They represent a general level of popularity of an origin or a destination. Latent variables γ_o and φ_d are used to represent the inverse effects of production and attraction respectively. For each origin o ($o \in \{1,2,...,S\}$), we can sample the inverse-production as follows:

$$\gamma_o \sim Gamma(a', b') \tag{5.3}$$

where a' and b' are the two hyper-parameters representing our prior belief about the shape and rate of the distribution of this inverse-production variable (i.e., what we think about $P(\varphi_o)$ before observing any data). Similarly, for each destination d ($d \in \{1,2,...,S\}$), we sample inverse-attraction as follows:

$$\varphi_d \sim Gamma(g', h') \tag{5.4}$$

where g' and h' are the two hyper-parameters representing our prior belief about the shape and rate of the distribution of this inverse-attraction variable.

Next, general potential production and attraction are respectively used to generate specific potential production for each origin and the specific potential attraction for each destination, which are both mapped to a joint latent space with K dimensions. For each origin o, the k-th ($k \in 1,2,...,K$) latent attribute of specific potential production is sampled as follows:

$$\rho_{ok} \sim Gamma(a, \gamma_o) \tag{5.5}$$

where a describes our prior belief about the shape of the distribution of specific potential production for dimension k. If the general potential production of an origin is higher, then the inverse-production variable γ_o is lower. If γ_o, as the rate parameter in this distribution, is lower, then the specific potential production ρ_{ok} is higher. Similarly, for each destination d, the k-th latent attribute of specific potential attraction is sampled as follows:

$$\omega_{dk} \sim Gamma(g, \varphi_d) \tag{5.6}$$

Consequently, each location has (i) an attribute of general potential production, (ii) an attribute of general potential attraction, (iii) K dimensions of specific potential production, and (iv) K dimensions of specific potential attraction.

The effect of generalized travel cost between two locations is another essential component in traditional gravity models but it is missing in the hierarchical Poisson factorization method. However, the flexibility of the hierarchical Poisson factorization allows incorporating this effect. We use the exponential cost function (de Dios Ortúzar & Willumsen, 2011) to describe a general effect of the impedance to travel for all OD pairs: $e^{-\beta C_{od}}$, where C_{od} represents the vector of travel costs between origin o and destination d, and $\boldsymbol{\beta}$ is the vector of cost parameters, which can be sampled from a Gamma distribution because the Gamma distribution is suitable for modeling non-negative values and we assume that β_x is positive:

$$\beta_x \sim Gamma(j,m) \tag{5.7}$$

The spatial interaction between origin o and destination d is modeled by the dot product of the vector of specific potential productions $\boldsymbol{\rho}_o$ and the vector of specific potential attractions $\boldsymbol{\omega}_d$. This is very similar to how potential production and attraction are multiplied in a traditional gravity model, except that only a single dimension is used. Then, the dot product for a certain OD pair is multiplied by the result of the exponential cost function of the generalized travel cost between this OD pair, obtaining the parameter of the Poisson distribution that models the spatial interaction. The spatial interaction between an OD pair can thus be modeled as follows:

$$F_{od} \sim Poisson((\boldsymbol{\rho}_o{}^T \boldsymbol{\omega}_d) \times e^{-\boldsymbol{\beta} \boldsymbol{C}_{od}}) \tag{5.8}$$

If there is only one pair of latent factors, i.e., K=1, this model would collapse into a traditional unconstrained gravity model for trip distribution, without a scaling factor (Sen, 1986). If there is no effect of generalized travel costs, i.e., $\beta = 0$, this model would collapse into the existing hierarchical Poisson factorization.

Figure 5.1 illustrates the approach for an OD matrix of S locations. For each location, there are K specific production or attraction factors. In theory, the number of dimensions of latent factors K should be smaller than the number of locations S (Theodoridis & Koutroumbas, 2009). It would be much more beneficial if a large-scale matrix is factorized, where the number of K parameters to be estimated would be much smaller than the number of cells in the matrix.

Figure 5.1 shows the flowchart of the model: how a spatial interaction matrix can be generated step by step. In the inference process, we need to estimate the latent factors and parameters (i.e., general and specific potential productions $\boldsymbol{\gamma}$ and $\boldsymbol{\rho}$, general and

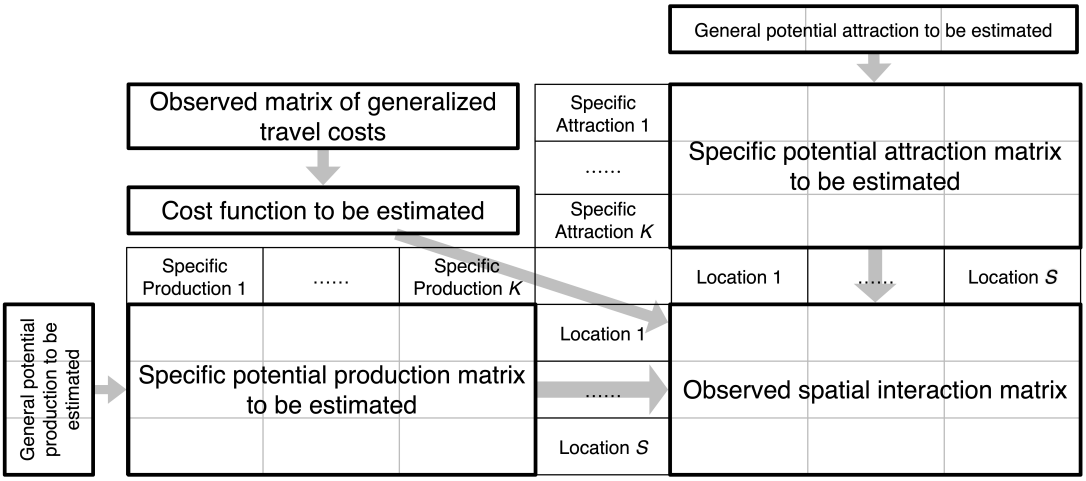


Figure 5.1: The flowchart of the model.

specific potential attractions $\boldsymbol{\varphi}$ and $\boldsymbol{\omega}$, and the parameter in the travel cost function $\boldsymbol{\beta}$) that can best explain the observed data (i.e., observed matrices of spatial interaction $\boldsymbol{F}$ and generalized travel costs $\boldsymbol{C}$). In the Bayesian approach, the objective is to estimate the distribution of the posterior probability $P(\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}|\boldsymbol{F},\boldsymbol{C})$. Bayes theorem is used to estimate the full distribution of the posterior probability:

$$P(\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}|\boldsymbol{F},\boldsymbol{C}) = P(\boldsymbol{F},\boldsymbol{C}|\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}) \cdot P(\boldsymbol{F},\boldsymbol{C}) \tag{5.9}$$

Compared to Equation 5.2, Equation 5.9 adds the parameter β in the travel cost function, and the observed generalized travel costs C. Also, Equation 5.9 adds $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$, which have hierarchical relationship with $\boldsymbol{\rho}$ and $\boldsymbol{\omega}$ respectively. The 2-stage hierarchical model can further decompose the full distribution of the posterior probability as follows:

$$\begin{aligned} P(\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}|\boldsymbol{F},\boldsymbol{C}) &= P(\boldsymbol{F},\boldsymbol{C}|\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}) \cdot P(\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{\varphi}) \cdot p(\boldsymbol{\gamma},\boldsymbol{\varphi})/P(\boldsymbol{F},\boldsymbol{C}) \\ &= p(\boldsymbol{F},\boldsymbol{C}|\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}) \cdot P(\boldsymbol{\rho}|\boldsymbol{\gamma}) \cdot P(\boldsymbol{\gamma}) \cdot P(\boldsymbol{\varphi}) \cdot P(\boldsymbol{\beta})/P(\boldsymbol{F},\boldsymbol{C}) \end{aligned} \tag{5.10}$$

In Equation 5.10, the likelihood part $P(\boldsymbol{F},\boldsymbol{C}|\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta})$ in the numerator is still tractable. The likelihood of the observed data for an OD pair given the corresponding latent variables can be calculated using the probability mass function of a Poisson distribution:

$$\begin{aligned} P(F_{od},C_{od}|\boldsymbol{\rho}_o,\boldsymbol{\omega}_d,\boldsymbol{\beta}) &= Poisson((\boldsymbol{\rho}_o{}^T\boldsymbol{\omega}_d) \times e^{-\boldsymbol{\beta}\boldsymbol{C}_{od}}) \\ &= ((\boldsymbol{\rho}_o{}^T\boldsymbol{\omega}_d) \times e^{-\boldsymbol{\beta}\boldsymbol{C}_{od}})^{F_{od}} \times e^{-((\boldsymbol{\rho}_o{}^T\boldsymbol{\omega}_d)\times e^{-\boldsymbol{\beta}\boldsymbol{C}_{od}})}/F_{od}! \end{aligned} \tag{5.11}$$

The hierarchical prior part $P(\boldsymbol{\rho}|\boldsymbol{\gamma}) \cdot P(\boldsymbol{\omega}|\boldsymbol{\varphi}) \cdot P(\boldsymbol{\gamma}) \cdot P(\boldsymbol{\varphi}) \cdot P(\boldsymbol{\beta})$ in the numerator is tractable as well based on Equations 5.3 5.7. However, it is impossible to determine

a closed-form expression describing denominator $P(\boldsymbol{F},\boldsymbol{C})$ since this probability has to be computed by integrating across all possible values of the latent variables. Instead of estimating the full distribution of the posterior probability $P(\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}\boldsymbol{F},\boldsymbol{C})$, one can opt for point estimation. More specifically, the MAP estimation finds the $\boldsymbol{\gamma}$, $\boldsymbol{\varphi}$, $\boldsymbol{\rho}$, $\boldsymbol{\omega}$, $\boldsymbol{\beta}$ that maximizes the posterior probability. This approach works because $\arg\max_{\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}} P(\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}|\boldsymbol{F},\boldsymbol{C})$ is equivalent to $\arg\max_{\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}} P(\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta},\boldsymbol{F},\boldsymbol{C})$, which is equal to $\arg\max_{\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}}(P(\boldsymbol{F},\boldsymbol{C}|\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta})\cdot P(\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}))$, independent from the denominator of Bayes theorem $P(\boldsymbol{F},\boldsymbol{C})$.

Because even the MAP is computationally demanding, to further ease the inference process, we use the Monte Carlo Expectation Maximization (MCEM) method (Booth & Hobert, 1999; Zhang et al., 2011), with Gibbs sampling for the expectation step and MAP for the maximization step since this method has been proven to fit well with hierarchical models (Booth et al., 2001). First the values of $\boldsymbol{\rho}$, $\boldsymbol{\omega}$ are initialized. Then in the expectation step, samples from $P(\boldsymbol{\gamma},\boldsymbol{\varphi}|\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta},\boldsymbol{F},\boldsymbol{C})$ are drawn using the Gibbs sampling technique. Since $\boldsymbol{\gamma}$ and $\boldsymbol{\varphi}$ are only related to $\boldsymbol{\rho}$ and $\boldsymbol{\omega}$ respectively, $P(\boldsymbol{\gamma},\boldsymbol{\varphi}|\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta},\boldsymbol{F},\boldsymbol{C})$ is essentially equal to $P(\boldsymbol{\gamma}|\boldsymbol{\rho})\cdot P(\boldsymbol{\varphi}|\boldsymbol{\omega})$. Finally, in the maximization step, the $\boldsymbol{\gamma}$, $\boldsymbol{\varphi}$ sampled in the last step are used to compute the MAP estimation of $\boldsymbol{\rho}$, $\boldsymbol{\omega}$, $\boldsymbol{\beta}$. The latest updated $\boldsymbol{\rho}$, $\boldsymbol{\omega}$, $\boldsymbol{\beta}$ are then input to the expectation step again. The expectation step and the maximization step are iterated until convergence, and afterwards $\arg\max_{\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}} P(\boldsymbol{\gamma},\boldsymbol{\varphi},\boldsymbol{\rho},\boldsymbol{\omega},\boldsymbol{\beta}|\boldsymbol{F},\boldsymbol{C})$ can be obtained.

5.4 Case study

The proposed method can be applied to any OD matrix that describes spatial interactions. In this paper, the approach is illustrated using an OD matrix constructed from the metro smart card data collected in Shanghai, China, on April 1, 2015. Provided by the Shanghai Open Data Applications (SODA) contest, the data contain the timestamps, and the station IDs of all tap-ins and tap-outs by all Shanghai metro smart card users. A specific algorithm was used (Wang et al., 2017) to detect the trips with a transfer between two lines requiring extra tap-outs and tap-ins, and filter out incomplete and unrealistic trips. Each tap-in can be paired with a tap-out, and define an associated trip. All trips on that day were aggregated to obtain the OD metro trip matrix. The value of each cell represents the number of metro trips from one station to another. At the time of data collection, the metro network of Shanghai consisted of 288 metro stations and 14 lines, as shown in Figure 5.2. Therefore, the spatial interactions are described by a 288×288 OD matrix excluding the diagonal. Besides, the shortest network distance and the minimum number of transfers between every two stations can be calculated, resulting in two 288×288 travel impedance matrices.

We randomly selected 60% of the cells of the OD matrix to train the model and estimate the latent factors $\boldsymbol{\rho}$, $\boldsymbol{\omega}$, $\boldsymbol{\beta}$, with the only restriction that all the stations should appear both as origin and destination in the training set, to guarantee that the production and

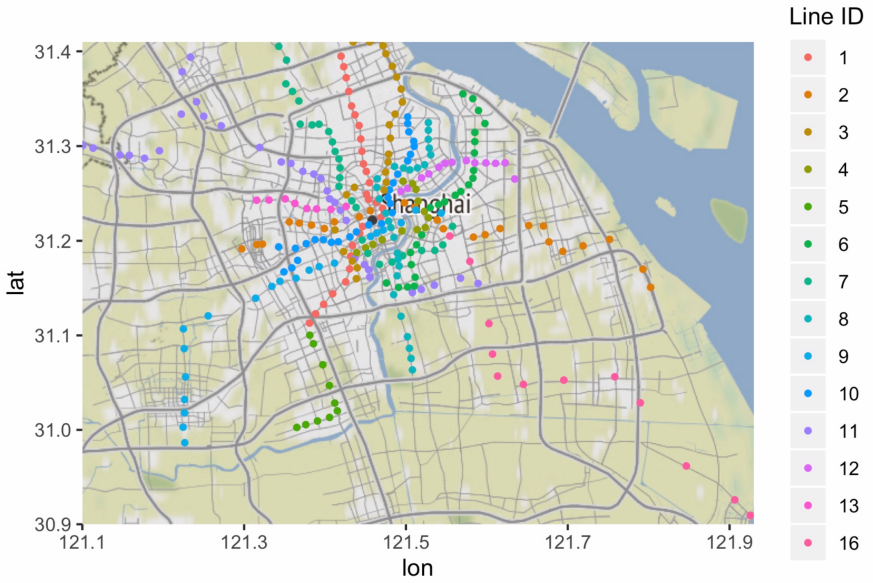


Figure 5.2: The 288 metro stations in the city of Shanghai, China (note that some stations share multiple lines, and in that case, one line is randomly selected to show its color).

attraction factors have been estimated for every station in the hold-out test set. The remaining 40% of the cells are used as the hold-out test set, for which we pretend not to know their actual values. Afterwards, for the hold-out test set, we can find the corresponding cell values of the travel impedance matrices, and use the estimated model to predict the number of metro trips from one station to another.

The hyper-parameters of the model, a, a', b', g, g', and h', are set to 0.3 as this setting has been proven to work well with different types of data in previous applications (Gopalan et al., 2015; Levitin et al., 2019). Hyper-parameters j and m are set to 0.3 as well since the resulting shape of the Gamma distribution is roughly consistent with our prior assumption about the travel cost function. The probabilistic modeling and the inference are implemented in Edward, a probabilistic programming language based on TensorFlow (Tran et al., 2016).

5.5 Results

In a first analysis, we test the number of specific production and attraction factor dimensions K: 1, 5, 10, 20, 30, 40, 50, 60, and 70. The number of dimensions K is naturally limited by the size of the training set. The predictive performance for the test set is shown in Figure 5.3. For an OD pair from the test set, the inferred $\boldsymbol{\rho}_o$, $\boldsymbol{\omega}_d$, $\boldsymbol{\beta}$ and the observed travel impedance $\boldsymbol{C}_{od}$ are used to calculate the expected value of the Poisson distribution given in Equation 5.8, which can serve as the predicted OD flow. Predictive performance is measured by the root-mean-squared error (RMSE) between the $log(x+1)$ transformations of observed and predicted OD flows.

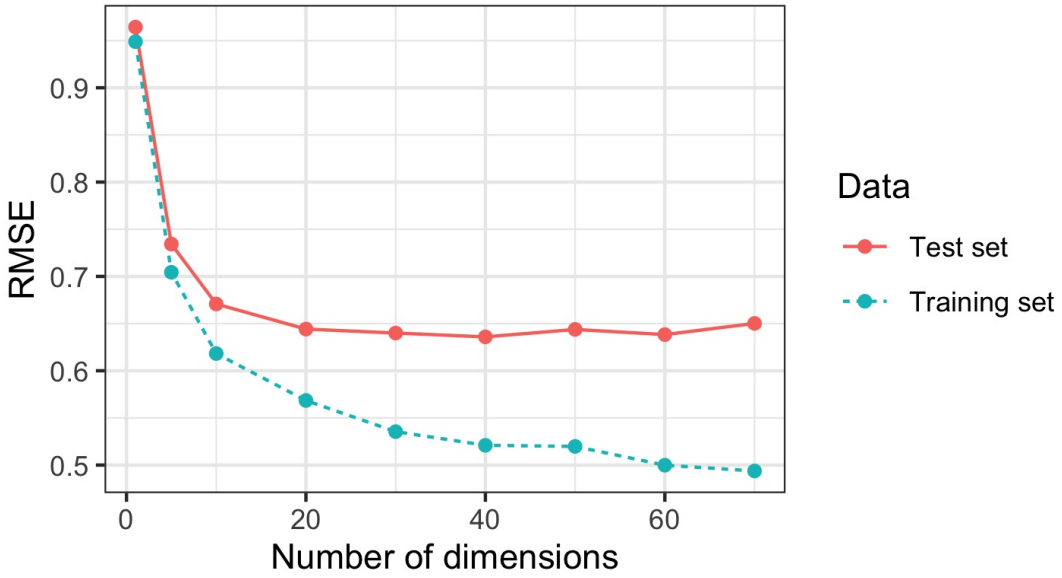


Figure 5.3: Root-mean-squared error of the models with a growing number of dimensions for the training and test sets.

It can be observed that in general, higher dimensionality for production and attraction factors improves the predictive performance of the model for the test set. It is not a surprise that the RMSE is always lower for the training set than for the test set, and this gap becomes larger with an increasing number of dimensions. The predictive performance for the test set improves sharply (i.e., 30% decrease in the RMSE) with an increase in the number of dimensions until it reaches 10. After that, the improvement dampens. From 10 dimensions to 20 dimensions, the RMSE decreases by 4%, and from 20 dimensions to the maximum, there is hardly any improvement. The model seems to benefit from adding more dimensions, but after a certain threshold, it enters an overfitting region. In summary, the model takes the advantage of multidimensionality, and the reduction of the RMSE can reach up to 34%.

A more detailed comparison of the predictive performance between different numbers of dimensions is presented in Figure 5.4, where, for better visualization, $log(x+1)$ transformation of the actual number of observations is rounded to the nearest 0.1, and for each of them, the minimum value, the 1st quartile value, the median value, the 3rd quartile value, and the maximum value of the predictions are calculated. The median value is indicated by a black point; the range from the 1st quartile value to the 3rd quartile value by a red error bar, and the range from the minimum value to the maximum value by a blue error bar.

It can be observed that with fewer dimensions, the model tends to underestimate the larger OD flows and overestimate the smaller OD flows. The prediction accuracies are relatively higher with more dimensions. The phenomenon of overfitting can also be observed: from 1 dimension to 10 dimensions, the predictive performance evolves very quickly, whilst there is almost no difference between the performance of the model

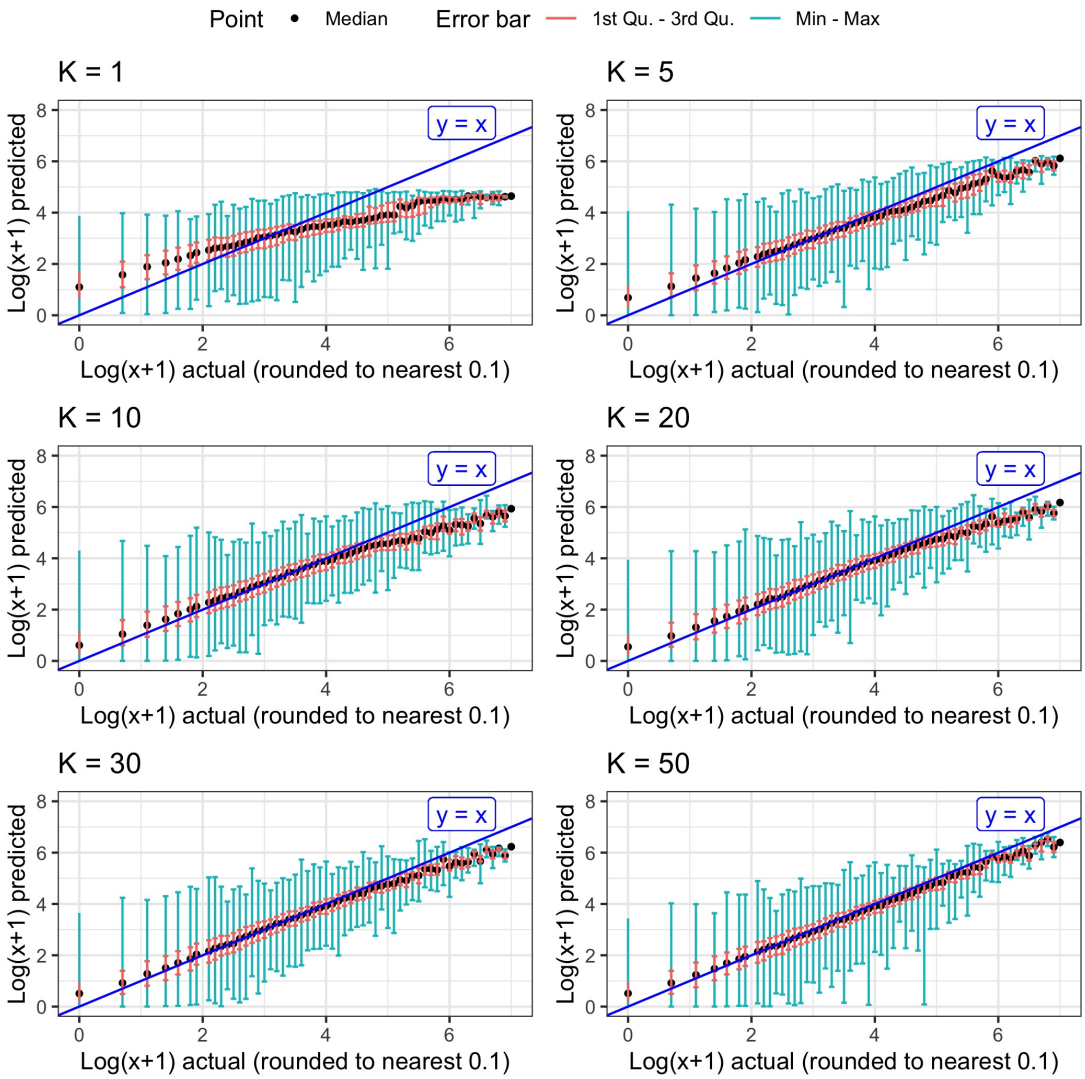


Figure 5.4: The prediction results of the model with a different number of dimensions for the test set.

with 20 dimensions and the one with 50 dimensions.

The comparison of predictive performance for the test set is shown in Figure 5.5 for the models with and without the travel cost function. Overall, in both situations, with an increasing number of dimensions, RMSE decreases fast at first and then gradually stabilizes. On average, considering the effect of travel impedance reduces the RMSE for the test set by 4.4%. The model with the travel cost function generally has a lower RMSE than the one without, except when the number of dimensions is 50 or 70.

Still, it can be observed in Figure 5.5 that with an increasing number of dimensions, the RMSE gap becomes much smaller. It seems that the model without the travel cost function can make up for its inherent deficiencies by adding more dimensions underlying production and attraction. To investigate this issue, we further analyze the estimated model without the travel cost function when the number of dimensions is 50. The cosine similarity between the specific production vector of a station and the

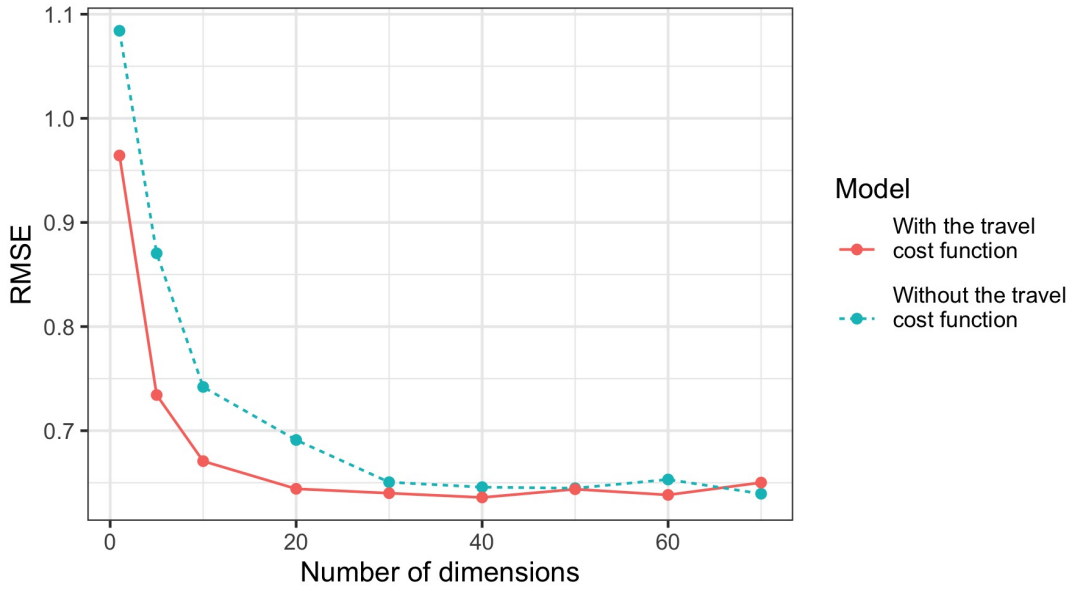


Figure 5.5: Root-mean-squared error for the test set with and without considering the effect of travel impedance.

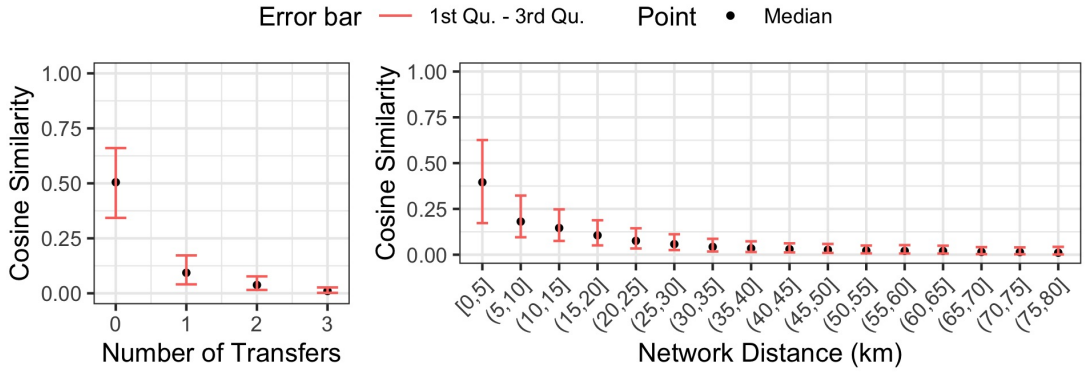


Figure 5.6: 1st quartile, median and 3rd quartile values of cosine similarities between the specific production vector of a station and the specific attraction vector of another station, estimated in the model without the travel cost function, over number of transfers and network distance between every two stations in the test set.

specific attraction vector of another station is calculated. A higher cosine similarity between these two vectors ($\boldsymbol{\rho}$ and $\boldsymbol{\omega}$) would directly lead to a higher dot product, given the magnitudes of them ($|\boldsymbol{\rho}|$ and $|\boldsymbol{\omega}|$), which are mostly determined by general production and attraction ($\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$). Figure 5.6 presents the 1st quartile value, the median value, and the 3rd quartile value of cosine similarities over the number of transfers and network distance between two stations.

The negative-exponential-alike curve of cosine similarity over the number of transfers or network distance in Figure 5.6 indicates that even without any prior information about travel costs, the 50 dimensions of specific production and attraction factors themselves, which are trained merely based on the interaction matrix, have already

been able to account for the effect of travel costs in a data-driven way. That somehow explains why the RMSE difference gradually diminishes in Figure 5.5. While the original Poisson factorization framework seems to sufficiently work for our test set, the newly proposed method still has an irreplaceable feature: since it involves a travel cost function, it allows predicting OD flows given different travel cost matrices.

5.6 Conclusions and recommendations

This study proposes an improved method for modeling trip generators and spatial interactions which combines the concepts of gravity model and a hierarchical Poisson factorization model. This method can be used to decompose an OD matrix and embed the locations into a continuous latent space. A traditional gravity model is extended to allow more dimensions underlying production and attraction, and the extended model is shown to be equivalent to an adapted framework of hierarchical Poisson factorization which additionally accounts for the effect of travel costs. The method was applied to one-day metro smart card data collected in Shanghai, China, which were aggregated to generate an OD trip matrix representing spatial interactions between stations. The results show that having more dimensions for potential productions and attractions improves the prediction of mobility patterns; however, the number of dimensions does not need to be too large. As more dimensions are added, the model tends to overfit the data. Moreover, the model considering the effect of travel impedance performs better than the model without this effect especially when the number of dimensions underlying production and attraction is small. More importantly, the proposed model allows predicting OD flows given different travel cost matrices.

We identify the following issues for further research. First, this study aggregates daily OD flows, leading to a relatively more symmetrical OD matrix. However, OD flows can also be disaggregated for different times of the day, yielding matrices that are very likely asymmetric. It is thus reasonable to assume that potential production or attraction of locations varies over time. Such dynamic patterns could be modeled using the framework of dynamic Poisson factorization (Charlin et al., 2015). Second, one might argue that a negative binomial distribution could better describe spatial interactions than a Poisson distribution does, because the mean and variance are not necessarily the same. Therefore, it is worth experimenting the negative binomial matrix factorization technique (Gouvert et al., 2018) to solve the problem. Third, while the proposed model can predict OD flows given new travel cost matrices between the existing stations, it is difficult to predict OD flows given a network with new stations. Essentially, all the production and attraction factors of the stations are estimated based on empirical data, which are not available for new stations. This so-called cold start problem (Su & Khoshgoftaar, 2009), which is common in recommendation systems (e.g., the preference of a new user for a new movie), is worth investigating regarding our problem as well.

5.7 Acknowledgment

We would like to express our gratitude to the Shanghai Open Data Apps (SODA) contest for making the data available for this research.

Chapter 6

Conclusions and recommendations

6.1 Conclusions

This thesis synthesizes four studies that process big mobility data to allow understanding of human spatial behavior, which is the outcome of people's activity type and location choice. Compared to traditional travel survey data, big mobility data are favorable for being cost-efficient, up to date, and promising especially in terms of the big sample size; however, big mobility data are, at the same time, limited in their very nature mostly due to privacy reasons. First, they lack sufficient features about travelers and trips (i.e., the data is thin). Second, the individual-level mobility data are sometimes aggregated into zone-to-zone trip tables. To overcome these limitations, this thesis posed the following main research question:

To what extent, and how, can big mobility data foster the understanding of human spatial behavior?

This thesis attempted two main strategies. Since big mobility data is too thin, the first strategy is to widen the thin data by adding explicit proxy variables to describe travelers, locations and trips. This strategy was tested in Chapter 2 and 3. Instead of adding explicit proxy variables, we can also leverage data-driven methods to implicitly capture the latent characteristics of travelers, locations and trips, by taking advantage of sample size. This strategy was tested in Chapter 4 and 5. The following two subsections explain how the two strategies work specifically and summarize to what extent they are effective based on the case study results.

6.1.1 Adding explicit proxy variables

A big mobility dataset usually only contains spatial-temporal information. It tracks a group of people's spatial-temporal traces either on the individual level or in aggregated form over a period within a mobility system. For example, Chapter 2 used smart card

data of the metro network in Shanghai, China, including the tap-in and tap-out records of each metro traveler for three months. Chapter 3 used mobile phone traces generated by the users in Shanghai who subscribe to a telecommunication company.

As the first case study of the thesis, Chapter 2 strictly narrowed the scope to study only the metro travelers' location choices for after-work activities, and we found that given a preselected activity purpose, building a location choice model using big mobility is not very different from building one using traditional travel survey data, except for the absence of personal attributes. In traditional travel survey data, the profile of a traveler is usually obtainable, such as one's socioeconomic level. Therefore, it is possible to model how different types of travelers would have different preferences in location choice. Such information is missing in big mobility data. In Chapter 2, we made a simple assumption: the characteristics of the area where one lives or works can reveal some characteristics of this person. We found each traveler's home and work locations based on the longitudinal travel data, and calculated the jobs-housing balance that can somehow reveal some characteristics of the areas of their home and work locations. This indicator then served as a proxy variable for each traveler's personal attributes and were added to the discrete choice model. It was found that this new model outperformed the one without any personal attributes.

Chapter 3 was based on the same strategy but a different path has been chosen. Instead of using mobility data to generate proxy variables, mobility data were combined with external data sources that can portrait the characteristics of travelers. Specifically, the case study used not only mobile phone traces to observe the movements of travelers but also the mobile internet usage data (i.e., frequency of visiting each type of sites) of the same users to portrait them. Besides, the case study proposed a clustering method to determine function types of urban areas using another external data source, urban point-of-interest data. As a result, we found that the mobile internet usage of travelers is statistically related to the types of areas that they prefer to visit.

In summary, adding explicit proxy variables can enhance the understanding of human spatial behavior. One can extract such proxy variables from big mobility data themselves. Intuitively, longitudinal behavioral data, because of the sample size, could contain some hidden information about travelers and/or trips, such as one's home and work locations. One can also obtain such proxy variables by fusing external data about travelers and locations, such as mobile internet usage patterns of travelers. Such information provides a novel perspective to understand different individuals' spatial behavior.

6.1.2 Using implicit data-driven methods

In the previous strategy, one has to arbitrarily define a proxy variable to account for behavioral heterogeneity. The effect is highly dependent on whether a good proxy variable is selected, and the selection of a good proxy variable is mostly dependent

on some assumptions based on prior knowledge. Without using any explicit proxy variables, Chapter 4 and 5 proposed to understand different people's human spatial behavior in a more data-driven way. A more flexible assumption underlies this approach: past behavior itself can reflect the heterogeneity in the population and serve as a reference to predict future behavior.

Chapter 4 implemented a neighborhood-based collaborative filtering algorithm to predict location choice for flexible activities using metro smart card data. This algorithm has widely been used in recommendation systems to predict user preferences for products. It looks for the so-called neighbors who share similar historical behavior patterns to an individual and then predicts this individual's preference based on the neighbors' historical behavior. Our case study showed that this method performed reasonably well in the context of travel behavior prediction as well.

As a continuation of the data-driven strategy, Chapter 5 proposed a matrix factorization model, another collaborative filtering method that has widely been used in recommendation systems. This new method can especially help understand aggregate spatial behavior, in terms of spatial interaction matrices. We pointed out the similarity between this model and a traditional gravity-based trip distribution model, and we showed the advantages of the new model based on the prediction performance.

In summary, using implicit data-driven methods can enhance the understanding of human spatial behavior. However, such understanding is similar to Google Translate's understanding of different languages. It is not explainable, and it is data-driven, in terms of nearest neighbors and latent factors.

6.2 Limitations and recommendations for future research

6.2.1 Data

The first limitation of this thesis is about the data. One might have questioned why this thesis did not try using both big mobility data and traditional travel survey data. In fact, it is difficult to find a place where both types of data are available perfectly aligned in space and time. All the case studies in this thesis used big mobility data from Shanghai, China. We had the chance to access those data because the urban authority started exploring the possibility of using big data to solve urban issues. Thus, they collaborated with the companies who owned those data, and together organized a series of urban data competitions. Researchers were invited to contribute with their ideas and thus obtained those data in return, which can later be used for their own academic purposes. On the other hand, the travel survey data of Shanghai are not open to researchers.

In Europe, the situation is the exact opposite. Traditional travel survey data are shared with the public in a very standardized way, such as OViN and MPN in the Netherlands

(Hoogendoorn-Lanser et al., 2015). However, especially after the implementation of the General Data Protection Regulation (GDPR), it has been extremely difficult for researchers to get access to any big mobility data, even in aggregated form.

As discussed previously, these two types of mobility data have both their advantages and disadvantages. In the future, if a researcher could, by any chance, access both of them in the same case study, we have the following recommendations. First, it would be valuable to compare them for the same modeling task. The marginal benefits of having a larger sample and having more features can be calculated. Second, methods to combine them can be developed to benefit from the strengths of both types of data.

6.2.2 Methodology

All the proposed methods can be used to understand the spatial behavior of the observed travelers, or the spatial interactions between the existing locations, from big mobility data. However, generalizing to the other unobserved travelers or locations is a different challenge. Such generalization is feasible using traditional travel survey data, and it is a necessary step especially when the estimated behavioral model should be applied to a synthetic population to predict the total travel demand.

Using big mobility data, such generalization is sometimes still needed. For example, it would be relevant to infer the spatial behavior of the total population from the users who subscribe to a telecommunication company. Even if we can observe the whole population of public transportation travelers in a city, there will still be new travelers in the transit system every day. How can we transfer our knowledge about the spatial behavior of the existing observed travelers to the new unobserved travelers?

There is a trade-off between the two main strategies of this thesis. Data-driven methods benefit from the flexibility to exploit the patterns revealed in big data. However, in this way, the knowledge about the observed travelers' spatial behavior cannot be transferred to any new unobserved travelers. For example, it is impossible to know the latent factors or nearest neighbors of a new unobserved traveler. Compared to the data-driven methods, adding explicit variables seems arbitrary, but the knowledge about spatial behavior of observed travelers can be transferred seamlessly, as long as the proxy variables can be generated for new unobserved travelers.

In the field of recommendation systems, researchers also noticed the same problem. The data-driven collaborative filtering methods cannot generate predictions for new users or products. This problem is designated as the cold start problem. To solve the cold start problem, researchers used explicit user and product attributes to help bridge the gap from existing users/products to new users/products (Schein et al., 2002).

The data-driven methods of modeling spatial behavior are worth exploring, but they cannot generate predictions for unobserved travelers or locations. To tackle this issue, we suggest that future researchers borrow the concept of the cold start problem from

recommendation systems. It is very likely that both strategies will have to be mixed: using the data-driven methods for the observed travelers, and at the same time, adding the explicit proxy variables to help bridge the gap from observed travelers/locations to unobserved travelers/locations.

6.3 Societal relevance and implications for practice

Human spatial behavior is relevant to many urban stakeholders. For example, transportation operating companies want to understand human spatial behavior because they want to provide the most requested services to fulfill people's travel demand. Retail companies want to understand human spatial behavior because they want to know where they should locate their business. Urban planners need to understand human spatial behavior because they want to plan cities in the most efficient way. Regardless of the final purposes, all the stakeholders should follow the steps defined in Figure 1.3 to leverage their mobility data: data collection, estimating, understanding, and decision making.

Traditionally, planning is a long-term iterative process including the aforementioned steps. For example, in Shanghai, the urban authority collects travel survey data as a basis for transportation modeling and planning every 5-10 years (Lu & Gu, 2011). If big mobility data are widely accepted as a reliable source for planning, the time and cost of data collection can be saved, and the iterative process can become much more frequent. Consequently, more timely decisions can be made. Shortening this cycle creates even more value for those mobility-related companies. Most of such companies have already been aware of this fact, and they have built their data pipelines to automatically collect their users' big mobility data. In this way, the marginal cost of data collection is extremely low. These companies are thus willing to quickly transform their updated and low-cost data to operational insights.

A set of useful methods were presented with four real-life case studies to show how they can help estimate and understand human spatial behavior using big mobility. They can further be used in practice to support decision making. Real estate investors and transportation planners can learn from Chapter 2 how to model where urban commuters choose to visit after work. Accordingly, real estate investors can predict the number of visitors in various investment scenarios and find out which area near a metro station has the most business potential. Transportation planners can use the model to design an improved metro network that provides maximum accessibility to all commuters for after-work activities. From Chapter 3, urban planners can learn to model different people's spatial preferences so that they can allocate urban functions in a geographically reasonable way. Retail companies can also utilize the model to decide where they should better locate their business. As discussed previously, Chapter 4 and 5 explored the data-driven methods of modeling human spatial behavior, and thus they cannot generate predictions for unobserved travelers or locations. This drawback makes them

limited in decision making unless the aforementioned cold start problem can be solved. However, they can still at least help transportation operating companies better estimate and understand the spatial behavior of their existing users, or the spatial interactions within their existing system. Transportation operating companies can use the method from Chapter 4 to predict the next stations that an existing user will visit, and they can use the method from Chapter 5 to calculate the travel demand elasticity with respect to generalized travel costs adjusted in the existing network.

Bibliography

Abbasi, A., T. H. Rashidi, M. Maghrebi, S. T. Waller (2015) Utilising location based social media in travel survey methods: bringing twitter data into the play, in: *Proceedings of the 8th ACM SIGSPATIAL international workshop on location-based social networks*, ACM, p. 1.

Aggarwal, C. C. (2016) Neighborhood-based collaborative filtering, in: *Recommender systems*, Springer, pp. 29–70.

Ahas, R., S. Silm, O. Järv, E. Saluveer, M. Tiru (2010) Using mobile positioning data to model locations meaningful to users of mobile phones, *Journal of urban technology*, 17(1), pp. 3–27.

Alexander, L., S. Jiang, M. Murga, M. C. González (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data, *Transportation research part c: emerging technologies*, 58, pp. 240–250.

Alsger, A. A., M. Mesbah, L. Ferreira, H. Safi (2015) Use of smart card fare data to estimate public transport origin–destination matrix, *Transportation Research Record*, 2535(1), pp. 88–96.

Altman, N. S. (1992) An introduction to kernel and nearest-neighbor nonparametric regression, *The American Statistician*, 46(3), pp. 175–185.

Anas, A. (1983) Discrete choice theory, information theory and the multinomial logit and gravity models, *Transportation Research Part B: Methodological*, 17(1), pp. 13–23.

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

Arai, A., A. Witayangkurn, H. Kanasugi, T. Horanont, X. Shao, R. Shibasaki (2014) Understanding user attributes from calling behavior: exploring call detail records through field observations, in: *Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia*, ACM, pp. 95–104.

Arentze, T., D. Ettema, H. Timmermans (2013) Location choice in the context of multi-day activity-travel patterns: model development and empirical results, *Transportmetrica A: Transport Science*, 9(2), pp. 107–123.

Arentze, T., H. Timmermans (2004) A learning-based transportation oriented simulation system, *Transportation Research Part B: Methodological*, 38(7), pp. 613–633.

Arentze, T., H. Timmermans (2007) Robust approach to modeling choice of locations in daily activity sequences, *Transportation Research Record*, 2003(1), pp. 59–63.

Axhausen, K. W., A. Zimmermann, S. Schönfelder, G. Rindsfüser, T. Haupt (2002) Observing the rhythms of daily life: A six-week travel diary, *Transportation*, 29(2), pp. 95–124.

Bagchi, M., P. R. White (2005) The potential of public transport smart card data, *Transport Policy*, 12(5), pp. 464–474.

Balcan, D., B. Gonçalves, H. Hu, J. J. Ramasco, V. Colizza, A. Vespignani (2010) Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model, *Journal of computational science*, 1(3), pp. 132–145.

Balmer, M., K. Meister, K. Nagel, K. Axhausen (2008) *Agent-based simulation of travel demand: Structure and computational performance of MATSim-T*, ETH, Eidgenössische Technische Hochschule Zürich, IVT Institut für .

Batty, M. (1976) *Urban modelling*, Cambridge University Press Cambridge.

Ben-Akiva, M., J. L. Bowman, D. Gopinath (1996) Travel demand model system for the information era, *Transportation*, 23(3), pp. 241–266.

Bhat, C. R., J. Guo (2004) A mixed spatially correlated logit model: formulation and application to residential choice modeling, *Transportation Research Part B: Methodological*, 38(2), pp. 147–168.

Bierlaire, M. (2003) Biogeme: a free package for the estimation of discrete choice models, in: *Swiss Transport Research Conference*, CONF.

Blondel, V. D., A. Decuyper, G. Krings (2015) A survey of results on mobile phone datasets analysis, *EPJ data science*, 4(1), p. 10.

Bonnel, P., E. Hombourger, A.-M. Olteanu-Raimond, Z. Smoreda (2015) Passive mobile phone dataset to construct origin-destination matrix: potentials and limitations, *Transportation Research Procedia*, 11, pp. 381–398.

Booth, J. G., J. P. Hobert (1999) Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1), pp. 265–285.

Booth, J. G., J. P. Hobert, W. Jank (2001) A survey of monte carlo algorithms for maximizing the likelihood of a two-stage hierarchical model, *Statistical Modelling*, 1(4), pp. 333–349.

Borgers, A., R. Van Der Heijden, H. Timmermans (1989) A variety seeking model of spatial choice-behaviour, *Environment and Planning A*, 21(8), pp. 1037–1048.

Bwambale, A., C. F. Choudhury, S. Hess (2017) Modelling trip generation using mobile phone data: A latent demographics approach, *Journal of Transport Geography*.

Bwambale, A., C. F. Choudhury, S. Hess (2019) Modelling departure time choice using mobile phone data, *Transportation research part A: policy and practice*, 130, pp. 424–439.

Caceres, N., L. M. Romero, F. G. Benitez (2013) Inferring origin–destination trip matrices from aggregate volumes on groups of links: a case study using volumes inferred from mobile phone data, *Journal of Advanced Transportation*, 47(7), pp. 650–666.

Calabrese, F., G. Di Lorenzo, L. Liu, C. Ratti (2011) Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in boston metropolitan area.

Calabrese, F., M. Diao, G. Di Lorenzo, J. Ferreira Jr, C. Ratti (2013) Understanding individual mobility patterns from urban sensing data: A mobile phone trace example, *Transportation research part C: emerging technologies*, 26, pp. 301–313.

Calabrese, F., L. Ferrari, V. D. Blondel (2014) Urban sensing using mobile phone network data: a survey of research, *Acm computing surveys (csur)*, 47(2), pp. 1–20.

Calabrese, F., L. Ferrari, V. D. Blondel (2015) Urban sensing using mobile phone network data: a survey of research, *Acm computing surveys (csur)*, 47(2), p. 25.

Castillo-Manzano, J. I., L. López-Valpuesta (2009) Urban retail fabric and the metro: A complex relationship. lessons from middle-sized spanish cities, *Cities*, 26(3), pp. 141–147.

Cats, O., Q. Wang, Y. Zhao (2015) Identification and classification of public transport activity centres in stockholm using passenger flows data, *Journal of Transport Geography*, 48, pp. 10–22.

Cesario, F. J. (1975) A combined trip generation and distribution model, *Transportation Science*, 9(3), pp. 211–223.

Chakirov, A., A. Erath (2012) Activity identification and primary location modelling based on smart card payment data for public transport, *Arbeitsberichte Verkehrs-und Raumplanung*, 786.

Charlin, L., R. Ranganath, J. McInerney, D. M. Blei (2015) Dynamic poisson factorization, in: *Proceedings of the 9th ACM Conference on Recommender Systems*, pp. 155–162.

Chen, C., J. Ma, Y. Susilo, Y. Liu, 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.

Cohen, P., R. Hahn, J. Hall, S. Levitt, R. Metcalfe (2016) Using big data to estimate consumer surplus: The case of uber, Tech. rep., National Bureau of Economic Research.

Çolak, S., L. P. Alexander, B. G. Alvim, S. R. Mehndiratta, M. C. González (2015) Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities, *Transportation Research Record*, 2526(1), pp. 126–135.

Collia, D. V., J. Sharp, L. Giesbrecht (2003) The 2001 national household travel survey: A look into the travel patterns of older americans, *Journal of safety research*, 34(4), pp. 461–470.

Danalet, A., L. Tinguely, M. de Lapparent, M. Bierlaire (2016) Location choice with longitudinal wifi data, *Journal of choice modelling*, 18, pp. 1–17.

Dashdorj, Z., L. Serafini, F. Antonelli, R. Larcher (2013) Semantic enrichment of mobile phone data records, in: *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*, ACM, p. 35.

de Dios Ortúzar, J., L. G. Willumsen (2011) *Modelling transport*, John wiley & sons.

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

De Vos, J., T. Schwanen, V. Van Acker, F. Witlox (2013) Travel and subjective well-being: A focus on findings, methods and future research needs, *Transport Reviews*, 33(4), pp. 421–442.

Deeva, G., J. De Smedt, J. De Weerdt, M. Óskarsdóttir (2019) Mining behavioural patterns in urban mobility sequences using foursquare check-in data from tokyo, in: *International Conference on Complex Networks and Their Applications*, Springer, pp. 931–943.

Demerouti, E., A. B. Bakker, S. A. Geurts, T. W. Taris (2009) Daily recovery from work-related effort during non-work time, in: *Current perspectives on job-stress recovery*, Emerald Group Publishing Limited, pp. 85–123.

Demissie, M. G., F. Antunes, C. Bento, S. Phithakkitnukoon, T. Sukhvibul (2016) Inferring origin-destination flows using mobile phone data: a case study of senegal, in: *2016 13th International conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON)*, IEEE, pp. 1–6.

Demissie, M. G., G. Correia, C. Bento (2015) Analysis of the pattern and intensity of urban activities through aggregate cellphone usage, *Transportmetrica A: transport science*, 11(6), pp. 502–524.

Demissie, M. G., G. H. de Almeida Correia, C. Bento (2013a) Exploring cellular network handover information for urban mobility analysis, *Journal of Transport Geography*, 31, pp. 164–170.

Demissie, M. G., G. H. de Almeida Correia, C. Bento (2013b) Intelligent road traffic status detection system through cellular networks handover information: An exploratory study, *Transportation research part C: emerging technologies*, 32, pp. 76–88.

Devillaine, F., M. Munizaga, M. Trépanier (2012) Detection of activities of public transport users by analyzing smart card data, *Transportation Research Record*, 2276(1), pp. 48–55.

Dong, X., M. E. Ben-Akiva, J. L. Bowman, J. L. Walker (2006) Moving from trip-based to activity-based measures of accessibility, *Transportation Research Part A: policy and practice*, 40(2), pp. 163–180.

Dunn, J. C. (1973) A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.

Everitt, B. S., S. Landau, M. Leese, D. Stahl (2011) Hierarchical clustering, *Cluster analysis*, 5.

Fotheringham, A. S., M. E. O'Kelly (1989) *Spatial interaction models: formulations and applications*, vol. 1, Kluwer Academic Publishers Dordrecht.

Fox, M. (1995) Transport planning and the human activity approach, *Journal of transport geography*, 3(2), pp. 105–116.

Furletti, B., P. Cintia, C. Renso, L. Spinsanti (2013) Inferring human activities from gps tracks, in: *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, ACM, p. 5.

Gan, Z., M. Yang, T. Feng, H. Timmermans (2020) Understanding urban mobility patterns from a spatiotemporal perspective: daily ridership profiles of metro stations, *Transportation*, 47(1), pp. 315–336.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin (2013) *Bayesian data analysis*, CRC press.

Geurs, K. T., B. Van Wee (2004) Accessibility evaluation of land-use and transport strategies: review and research directions, *Journal of Transport geography*, 12(2), pp. 127–140.

Giannotti, F., D. Pedreschi (2008) Mobility, data mining and privacy: A vision of convergence, in: *Mobility, data mining and privacy*, Springer, pp. 1–11.

Giuliano, G., H.-H. Hu, K. Lee (2003) Travel patterns of the elderly: The role of land use, Tech. rep., METRANS Transportation Center.

Goh, S., K. Lee, J. S. Park, M. Choi (2012) Modification of the gravity model and application to the metropolitan seoul subway system, *Physical Review E*, 86(2), p. 026102.

Gong, V. X., W. Daamen, A. Bozzon, S. P. Hoogendoorn (2019) Estimate sentiment of crowds from social media during city events, *Transportation research record*, 2673(11), pp. 836–850.

Gong, V. X., J. Yang, W. Daamen, A. Bozzon, S. Hoogendoorn, G.-J. Houben (2018) Using social media for attendees density estimation in city-scale events, *IEEE Access*, 6, pp. 36325–36340.

Gopalan, P., J. M. Hofman, D. M. Blei (2015) Scalable recommendation with hierarchical poisson factorization., in: *UAI*, pp. 326–335.

Goulet-Langlois, G., H. N. Koutsopoulos, J. Zhao (2016) Inferring patterns in the multi-week activity sequences of public transport users, *Transportation Research Part C: Emerging Technologies*, 64, pp. 1–16.

Goulet-Langlois, G., H. N. Koutsopoulos, Z. Zhao, J. Zhao (2017) Measuring regularity of individual travel patterns, *IEEE Transactions on Intelligent Transportation Systems*, 19(5), pp. 1583–1592.

Gouvert, O., T. Oberlin, C. Févotte (2018) Negative binomial matrix factorization for recommender systems, *arXiv preprint arXiv:1801.01708*.

Griffith, D. A., M. M. Fischer (2013) Constrained variants of the gravity model and spatial dependence: model specification and estimation issues, *Journal of Geographical Systems*, 15(3), pp. 291–317.

Hansen, W. G. (1959) How accessibility shapes land use, *Journal of the American Institute of planners*, 25(2), pp. 73–76.

Hasan, S., C. M. Schneider, S. V. Ukkusuri, M. C. González (2013a) Spatiotemporal patterns of urban human mobility, *Journal of Statistical Physics*, 151(1-2), pp. 304–318.

Hasan, S., S. V. Ukkusuri (2014) Urban activity pattern classification using topic models from online geo-location data, *Transportation Research Part C: Emerging Technologies*, 44, pp. 363–381.

Hasan, S., S. V. Ukkusuri (2015) Location contexts of user check-ins to model urban geo life-style patterns, *PloS one*, 10(5), p. e0124819.

Hasan, S., X. Zhan, S. V. Ukkusuri (2013b) Understanding urban human activity and mobility patterns using large-scale location-based data from online social media, in: *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, pp. 1–8.

Hastie, T., R. Tibshirani, J. Friedman (2009) Model assessment and selection, in: *The elements of statistical learning*, Springer, pp. 219–259.

He, Z., L. Zheng, W. Guan (2015) A simple nonparametric car-following model driven by field data, *Transportation Research Part B: Methodological*, 80, pp. 185–201.

Hoogendoorn-Lanser, S., N. T. Schaap, M.-J. OldeKalter (2015) The netherlands mobility panel: An innovative design approach for web-based longitudinal travel data collection, *Transportation Research Procedia*, 11, pp. 311–329.

Horni, A. (2013) *Destination choice modeling of discretionary activities in transport microsimulations*, Ph.D. thesis, ETH Zurich.

Horni, A., D. M. Scott, M. Balmer, K. W. Axhausen (2009) Location choice modeling for shopping and leisure activities with matsim: combining microsimulation and time geography, *Transportation Research Record*, 2135(1), pp. 87–95.

Huang, A., L. Gallegos, K. Lerman (2017) Travel analytics: Understanding how destination choice and business clusters are connected based on social media data, *Transportation Research Part C: Emerging Technologies*, 77, pp. 245–256.

Huang, A., D. Levinson (2015) Axis of travel: Modeling non-work destination choice with gps data, *Transportation Research Part C: Emerging Technologies*, 58, pp. 208–223.

Huang, Q., D. W. Wong (2016) Activity patterns, socioeconomic status and urban spatial structure: what can social media data tell us?, *International Journal of Geographical Information Science*, 30(9), pp. 1873–1898.

Hyman, G. M. (1969) The calibration of trip distribution models, *Environment and Planning A*, 1(1), pp. 105–112.

Ibrahim, M. F., P. J. McGoldrick (2017) *Shopping choices with public transport options: an agenda for the 21st century*, Routledge.

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

Isaacman, S., R. Becker, R. Cáceres, S. Kobourov, M. Martonosi, J. Rowland, A. Varshavsky (2011) Identifying important places in peoples lives from cellular network data, in: *International Conference on Pervasive Computing*, Springer, pp. 133–151.

Järv, O., R. Ahas, F. Witlox (2014) Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records, *Transportation Research Part C: Emerging Technologies*, 38, pp. 122–135.

Jiang, S., A. Alves, F. Rodrigues, J. Ferreira Jr, F. C. Pereira (2015) Mining point-of-interest data from social networks for urban land use classification and disaggregation, *Computers, Environment and Urban Systems*, 53, pp. 36–46.

Jiang, S., G. A. Fiore, Y. Yang, J. Ferreira Jr, E. Frazzoli, M. C. González (2013) A review of urban computing for mobile phone traces: current methods, challenges and opportunities, in: *Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing*, pp. 1–9.

Kassoff, H., H. D. Deutschman (1969) Trip generation: a critical appraisal, *Highway Research Record*, (297).

Koppelman, F. S. (2007) Closed form discrete choice models, in: *Handbook of transport modelling*, Emerald Group Publishing Limited.

Koppelman, F. S., C.-H. Wen (2000) The paired combinatorial logit model: properties, estimation and application, *Transportation Research Part B: Methodological*, 34(2), pp. 75–89.

Koren, Y., R. Bell, C. Volinsky (2009) Matrix factorization techniques for recommender systems, *Computer*, 42(8), pp. 30–37.

Kuhlman, W. (2015) *The construction of purpose-specific OD matrices using public transport smart card data*, Ph.D. thesis, TU Delft.

Kwan, M.-P. (2016) Algorithmic geographies: Big data, algorithmic uncertainty, and the production of geographic knowledge, *Annals of the American Association of Geographers*, 106(2), pp. 274–282.

Levitin, H. M., J. Yuan, Y. L. Cheng, F. J. Ruiz, E. C. Bush, J. N. Bruce, P. Canoll, A. Iavarone, A. Lasorella, D. M. Blei, et al. (2019) De novo gene signature identification from single-cell rna-seq with hierarchical poisson factorization, *Molecular systems biology*, 15(2), p. e8557.

Lin, X., M. Li, F. He (2018) Nonlinear pricing in linear cities with elastic demands, *Transportation Research Part C: Emerging Technologies*, 95, pp. 616–635.

Liu, F., D. Janssens, G. Wets, M. Cools (2013) Annotating mobile phone location data with activity purposes using machine learning algorithms, *Expert Systems with Applications*, 40(8), pp. 3299–3311.

Liu, L., A. Hou, A. Biderman, C. Ratti, J. Chen (2009) Understanding individual and collective mobility patterns from smart card records: A case study in shenzhen, in: *2009 12th International IEEE Conference on Intelligent Transportation Systems*, IEEE, pp. 1–6.

Liu, Y., Z. Sui, C. Kang, Y. Gao (2014) Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data, *PloS one*, 9(1), p. e86026.

Long, Y., J.-C. Thill (2015) Combining smart card data and household travel survey to analyze jobs–housing relationships in beijing, *Computers, Environment and Urban Systems*, 53, pp. 19–35.

Long, Y., Y. Zhang, C. Cui (2012) Analysing jobs-housing relationship and commuting patterns of beijing using bus smart card data (in chinese), *Acta Geographica Sinica*, 67(10), pp. 1339–1352.

Lu, X.-m., X.-t. Gu (2011) The fifth travel survey of residents in shanghai and characteristics analysis, *Urban Transport of China*, 9(5), pp. 1–7.

Luo, D., O. Cats, H. van Lint, G. Currie (2019) Integrating network science and public transport accessibility analysis for comparative assessment, *Journal of Transport Geography*, 80, p. 102505.

Ma, X., C. Liu, H. Wen, Y. Wang, Y.-J. Wu (2017) Understanding commuting patterns using transit smart card data, *Journal of Transport Geography*, 58, pp. 135–145.

Ma, X., Y.-J. Wu, Y. Wang, F. Chen, J. Liu (2013) Mining smart card data for transit riders travel patterns, *Transportation Research Part C: Emerging Technologies*, 36, pp. 1–12.

Marchal, F., K. Nagel (2005) Modeling location choice of secondary activities with a social network of cooperative agents, *Transportation Research Record*, 1935(1), pp. 141–146.

Morency, C., M. Trepanier, B. Agard (2007) Measuring transit use variability with smart-card data, *Transport Policy*, 14(3), pp. 193–203.

Munizaga, M. A., C. Palma (2012) Estimation of a disaggregate multimodal public transport origin–destination matrix from passive smartcard data from santiago, chile, *Transportation Research Part C: Emerging Technologies*, 24, pp. 9–18.

Murphy, K. P. (2012) *Machine learning: a probabilistic perspective*, MIT press.

Nakamura, K., F. Gu, V. Wasuntarasook, V. Vichiensan, Y. Hayashi (2016) Failure of transit-oriented development in bangkok from a quality of life perspective, *Asian Transport Studies*, 4(1), pp. 194–209.

Nanni, M., R. Trasarti, B. Furletti, L. Gabrielli, P. Van Der Mede, J. De Bruijn, E. De Romph, G. Bruil (2013) Transportation planning based on gsm traces: a case study on ivory coast, in: *International Workshop on Citizen in Sensor Networks*, Springer, pp. 15–25.

Ni, L., X. C. Wang, X. M. Chen (2018) A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data, *Transportation research part C: emerging technologies*, 86, pp. 510–526.

Noulas, A., S. Scellato, N. Lathia, C. Mascolo (2012) Mining user mobility features for next place prediction in location-based services, in: *2012 IEEE 12th international conference on data mining*, IEEE, pp. 1038–1043.

Pappalardo, L., D. Pedreschi, Z. Smoreda, F. Giannotti (2015) Using big data to study the link between human mobility and socio-economic development, in: *2015 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 871–878.

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

Phithakkitnukoon, S., T. Horanont, G. Di Lorenzo, R. Shibasaki, C. Ratti (2010) Activity-aware map: Identifying human daily activity pattern using mobile phone data, in: *International Workshop on Human Behavior Understanding*, Springer, pp. 14–25.

Rashidi, T. H., A. Abbasi, M. Maghrebi, S. Hasan, 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.

Rasouli, S., H. Timmermans (2014) Activity-based models of travel demand: promises, progress and prospects, *International Journal of Urban Sciences*, 18(1), pp. 31–60.

Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom, J. Riedl (1994) Grouplens: an open architecture for collaborative filtering of netnews, in: *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, ACM, pp. 175–186.

Richmond, S. (2012) Smartphones hardly used for calls, *The Telegraph*, 29.

Rupert Jr, G., et al. (2012) *Simultaneous statistical inference*, Springer Science & Business Media.

Schafer, J. B., D. Frankowski, J. Herlocker, S. Sen (2007) Collaborative filtering recommender systems, in: *The adaptive web*, Springer, pp. 291–324.

Schein, A. I., A. Popescul, L. H. Ungar, D. M. Pennock (2002) Methods and metrics for cold-start recommendations, in: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 253–260.

Schlich, R., K. W. Axhausen (2003) Habitual travel behaviour: evidence from a six-week travel diary, *Transportation*, 30(1), pp. 13–36.

Schönfelder, S., K. W. Axhausen (2003) Activity spaces: measures of social exclusion?, *Transport policy*, 10(4), pp. 273–286.

Sen, A. (1986) Maximum likelihood estimation of gravity model parameters, *Journal of Regional Science*, 26(3), pp. 461–474.

Seneviratne, S., A. Seneviratne, P. Mohapatra, A. Mahanti (2014) Predicting user traits from a snapshot of apps installed on a smartphone, *ACM SIGMOBILE Mobile Computing and Communications Review*, 18(2), pp. 1–8.

Seneviratne, S., A. Seneviratne, P. Mohapatra, A. Mahanti (2015) Your installed apps reveal your gender and more!, *ACM SIGMOBILE Mobile Computing and Communications Review*, 18(3), pp. 55–61.

Sermons, M. W., F. S. Koppelman (2001) Representing the differences between female and male commute behavior in residential location choice models, *Journal of transport geography*, 9(2), pp. 101–110.

Sevtsuk, A., C. Ratti (2010) Does urban mobility have a daily routine? learning from the aggregate data of mobile networks, *Journal of Urban Technology*, 17(1), pp. 41–60.

Silm, S., R. Ahas (2014) Ethnic differences in activity spaces: A study of out-of-home nonemployment activities with mobile phone data, *Annals of the Association of American Geographers*, 104(3), pp. 542–559.

Silm, S., R. Ahas, V. Mooses (2018) Are younger age groups less segregated? measuring ethnic segregation in activity spaces using mobile phone data, *Journal of Ethnic and Migration Studies*, 44(11), pp. 1797–1817.

Sivakumar, A., C. R. Bhat (2007) Comprehensive, unified framework for analyzing spatial location choice, *Transportation Research Record*, 2003(1), pp. 103–111.

Su, X., T. M. Khoshgoftaar (2009) A survey of collaborative filtering techniques, *Advances in artificial intelligence*, 2009.

Sun, L., Y. Lu, J. G. Jin, D.-H. Lee, K. W. Axhausen (2015) An integrated bayesian approach for passenger flow assignment in metro networks, *Transportation Research Part C: Emerging Technologies*, 52, pp. 116–131.

Suriñach, J., J. Romaní, V. Royuela, M. Reyes (2000) Urban systems in the barcelona province: A first step for estimating local economic activity.

Tao, S., D. Rohde, J. Corcoran (2014) Examining the spatial–temporal dynamics of bus passenger travel behaviour using smart card data and the flow-comap, *Journal of Transport Geography*, 41, pp. 21–36.

Theodoridis, S., K. Koutroumbas (2009) *Pattern Recognition*, Academic Press.

Timmermans, H. (1993) Retail environments and spatial shopping behavior, in: *Advances in psychology*, vol. 96, Elsevier, pp. 342–377.

Timmermans, H., A. Borgers, J. van Dijk, H. Oppewal (1992) Residential choice behaviour of dual earner households: a decompositional joint choice model, *Environment and Planning A*, 24(4), pp. 517–533.

Tolouei, R., S. Psarras, R. Prince (2017) Origin-destination trip matrix development: Conventional methods versus mobile phone data, *Transportation research procedia*, 26, pp. 39–52.

Toole, J. L., S. Colak, B. Sturt, L. P. Alexander, A. Evsukoff, M. C. González (2015) The path most traveled: Travel demand estimation using big data resources, *Transportation Research Part C: Emerging Technologies*, 58, pp. 162–177.

Train, K. E. (2009) *Discrete choice methods with simulation*, Cambridge university press.

Tran, D., A. Kucukelbir, A. B. Dieng, M. Rudolph, D. Liang, D. M. Blei (2016) Edward: A library for probabilistic modeling, inference, and criticism, *arXiv preprint arXiv:1610.09787*.

Trépanier, M., N. Tranchant, R. Chapleau (2007) Individual trip destination estimation in a transit smart card automated fare collection system, *Journal of Intelligent Transportation Systems*, 11(1), pp. 1–14.

Tu, W., R. Cao, Y. Yue, B. Zhou, Q. Li, Q. Li (2018) Spatial variations in urban public ridership derived from gps trajectories and smart card data, *Journal of Transport Geography*, 69, pp. 45–57.

Van Wee, B. (2009) Self-selection: a key to a better understanding of location choices, travel behaviour and transport externalities?, *Transport reviews*, 29(3), pp. 279–292.

Van Wee, B., H. Holwerda, R. Van Baren (2002) Preferences for modes, residential location and travel behaviour: the relevance for land-use impacts on mobility, *European Journal of Transport and Infrastructure Research*, 2(3/4), pp. 305–316.

Wang, F., C. Chen (2018) On data processing required to derive mobility patterns from passively-generated mobile phone data, *Transportation Research Part C: Emerging Technologies*, 87, pp. 58–74.

Wang, Y., G. H. de Almeida Correia, E. de Romph, H. Timmermans (2017) Using metro smart card data to model location choice of after-work activities: An application to shanghai, *Journal of Transport Geography*, 63, pp. 40–47.

Wang, Y., G. H. de Almeida Correia, B. van Arem (2019) Relationships between mobile phone usage and activity-travel behavior: A review of the literature and an example, in: *Advances in Transport Policy and Planning*, vol. 3, Elsevier, pp. 81–105.

Wang, Y., G. H. de Almeida Correia, B. van Arem, H. Timmermans (2018) Understanding travellers preferences for different types of trip destination based on mobile internet usage data, *Transportation Research Part C: Emerging Technologies*, 90, pp. 247–259.

Wang, Y., R. Kutadinata, S. Winter (2016) Activity-based ridesharing: increasing flexibility by time geography, in: *Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 1–10.

Wen, C.-H., F. S. Koppelman (2000) A conceptual and methdological framework for the generation of activity-travel patterns, *Transportation*, 27(1), pp. 5–23.

Willigers, J., B. van Wee (2011) High-speed rail and office location choices. a stated choice experiment for the netherlands, *Journal of Transport Geography*, 19(4), pp. 745–754.

Wolf, J. (2006) Applications of new technologies in travel surveys, in: *Travel survey methods: Quality and future directions*, Emerald Group Publishing Limited, pp. 531–544.

Wolf, J., R. Guensler, W. Bachman (2001) Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data, *Transportation Research Record*, 1768(1), pp. 125–134.

Wolf, J., S. Schönfelder, U. Samaga, M. Oliveira, K. W. Axhausen (2004) Eighty weeks of global positioning system traces: approaches to enriching trip information, *Transportation Research Record*, 1870(1), pp. 46–54.

Xue, G.-R., C. Lin, Q. Yang, W. Xi, H.-J. Zeng, Y. Yu, Z. Chen (2005) Scalable collaborative filtering using cluster-based smoothing, in: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 114–121.

Yin, M., M. Sheehan, S. Feygin, J.-F. Paiement, A. Pozdnoukhov (2017) A generative model of urban activities from cellular data, *IEEE Transactions on Intelligent Transportation Systems*, 19(6), pp. 1682–1696.

Yoon, S. Y., K. Deutsch, Y. Chen, K. G. Goulias (2012) Feasibility of using time–space prism to represent available opportunities and choice sets for destination choice models in the context of dynamic urban environments, *Transportation*, 39(4), pp. 807–823.

Yuan, J., Y. Zheng, X. Xie (2012) Discovering regions of different functions in a city using human mobility and pois, in: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 186–194.

Yue, Y., T. Lan, A. G. Yeh, 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.

Zhang, L., D. Agarwal, B.-C. Chen (2011) Generalizing matrix factorization through flexible regression priors, in: *Proceedings of the fifth ACM conference on Recommender systems*, pp. 13–20.

Zhao, J., A. Rahbee, N. H. Wilson (2007) Estimating a rail passenger trip origin-destination matrix using automatic data collection systems, *Computer-Aided Civil and Infrastructure Engineering*, 22(5), pp. 376–387.

Zhao, K., S. Tarkoma, S. Liu, H. Vo (2016a) Urban human mobility data mining: An overview, in: *2016 IEEE International Conference on Big Data (Big Data)*, IEEE, pp. 1911–1920.

Zhao, S., K. Zhang (2017) Observing individual dynamic choices of activity chains from location-based crowdsourced data, *Transportation Research Part C: Emerging Technologies*, 85, pp. 1–22.

Zhao, Z., H. N. Koutsopoulos, J. Zhao (2018) Individual mobility prediction using transit smart card data, *Transportation research part C: emerging technologies*, 89, pp. 19–34.

Zhao, Z., S.-L. Shaw, Y. Xu, F. Lu, J. Chen, L. Yin (2016b) Understanding the bias of call detail records in human mobility research, *International Journal of Geographical Information Science*, 30(9), pp. 1738–1762.

Zheng, Y., L. Zhang, X. Xie, W.-Y. Ma (2009) Mining interesting locations and travel sequences from gps trajectories, in: *Proceedings of the 18th international conference on World wide web*, ACM, pp. 791–800.

Zhou, J., E. Murphy, Y. Long (2014) Commuting efficiency in the beijing metropolitan area: An exploration combining smartcard and travel survey data, *Journal of Transport Geography*, 41, pp. 175–183.

Ziemke, D., I. Kaddoura, K. Nagel (2019) The matsim open berlin scenario: A multimodal agent-based transport simulation scenario based on synthetic demand modeling and open data, *Procedia computer science*, 151, pp. 870–877.

Summary

People are engaged in a variety of activities through space every day. The choice of type and location of activities is known as human spatial behavior. Urban decision makers need to understand how land use and transportation systems can shape human spatial behavior in order to design better systems. In the past, they have collected mobility data through travel surveys to understand human spatial behavior. Today, a wide range of automatically collected data have become available as alternative data sources.

Big mobility data vs. traditional travel survey data has been a topic of long-time debate in human mobility and travel behavior research. Big data are intuitively better but this is not always the case. Big mobility data relate to a large number of travelers and trips but little is known about each individual individual traveler and trip, not to mention that sometimes their information has to be aggregated for privacy concerns. On the other hand, travel survey data, despite reporting only a small group of respondents, tend to include abundant features about each individual traveler, such as age and attitudes, and each trip, such as trip purpose. Assuming that each row represents one traveler and each column represents one feature, big mobility data should have been described as long and thin, and "small" survey data (Chen et al., 2016) as short and wide.

Still, big mobility data are favorable for being cost-efficient, up-to-date and big in terms of sample size. Therefore, given all limitations and advantages of big mobility data, this thesis aims to answer the following research question:

To what extent, and how, can big mobility data foster the understanding of human spatial behavior?

The major contribution of this thesis consists of two main strategies adopted to answer the research question. The first main strategy is to make the long and thin data wider. Since the lack of features is the biggest obstacle for big mobility data to explain human spatial behavior, attempts were made to generate proxy variables for traveler segmentation and trip characterization, from either big mobility data themselves (Chapter 2) or from external datasets (Chapter 3).

Given a preselected activity purpose, building a location choice model using big mobility data is not very different from building a location choice model using traditional travel survey data, except for the absence of personal attributes. In Chapter 2, we made

a simple assumption: the characteristics of the area where a person lives or works can reveal particular characteristics of this person. We found each travelers home and work locations from longitudinal travel data, and used the job-housing balance to reveal some characteristics of these areas. This indicator then serves as a proxy variable for each travelers personal characteristics and was added to a discrete choice model. It was found that this new model outperformed the model without any personal characteristics.

Chapter 3 was based on the same strategy but a different path was chosen. Instead of merely using mobility data to generate proxy variables, mobility data were combined with external data sources that portrayed the characteristics of travelers. Specifically, the case study used not only mobile phone traces to observe the movements of travelers, but also mobile internet usage data (i.e., frequency of visiting each type of sites via smart phones) of the same users. Besides, the case study proposed a clustering method to determine types of functional urban areas using another external data source, urban point-of-interest data. As a result, we found that the mobile internet usage of travelers is statistically related to the types of areas that they prefer to visit.

The second main strategy takes a relatively more groundbreaking approach, inspired by the collaborative filtering algorithms that are commonly used to model user preferences in recommendation systems. Without using any specific proxy variables, Chapter 4 and 5 implemented data-driven methods, relying only on empirical observations about many people, and not requiring imposing any theory-based prior assumptions about the mechanisms of human spatial behavior. Intuitively, this approach might work because historical spatial behavior itself can indicate heterogeneity between individuals within a given group of travelers and thus help make predictions about their future behavior.

Chapter 4 implemented a neighborhood-based collaborative filtering algorithm to predict location choice for flexible activities using metro smart card data. This algorithm has widely been used in recommendation systems to predict user preferences for products. It looks for so-called “neighbors” that share similar historical behavior patterns with an individual and then predicts individuals preference based on the neighbors historical behavior. Our case study showed that this method performed reasonably well in the context of travel behavior prediction, suggesting further exploration of this method, which is still relatively unfamiliar to most transportation researchers.

As a continuation of the data-driven strategy, Chapter 5 proposed a matrix factorization model, another collaborative filtering method that has widely been used in recommendation systems. This new method can especially help understanding aggregate spatial behavior in terms of spatial interaction matrices. We pointed out that this model is almost equivalent to a traditional gravity-based trip distribution model, but more flexible. In a case study of predicting origin-destination trip matrices of a metro network, we demonstrated the advantages of the new model in terms of its prediction performance.

Overall, there is a trade-off between the two proposed strategies. While the data-driven approaches benefit from the flexibility to exploit patterns revealed in big data,

they are limited in their interpretability as well as their ability to generalize the discovered patterns from observed travelers or locations to unobserved travelers or locations. Nevertheless, this type of method can be sufficiently useful for transportation operating companies to capture day-to-day behavioral patterns of their existing customers for daily operations. On the other hand, the first main strategy about adding proxy variables is less data-driven, but still better for longer-term planning and policy-making because it focuses on making predictions in what-if scenarios involving unobserved travelers or locations.

Samenvatting

Mensen zijn elke dag bezig met verschillende activiteiten op verschillende locaties. De keuze van type en locatie van activiteiten staat bekend als menselijk ruimtelijk gedrag. Stedelijke beleidsmakers willen begrijpen hoe landgebruik en transportsystemen het menselijk ruimtelijk gedrag kunnen beïnvloeden zodat ze landgebruik en transportsystemen beter kunnen plannen. In het verleden verzamelden zij mobiliteitsgegevens via reis-enquêtes om het ruimtelijk gedrag van mensen te begrijpen. Tegenwoordig is een breed scala aan automatisch verzamelde gegevens beschikbaar gekomen als alternatieve gegevensbronnen.

Big mobility data versus traditionele reis-enquête data is al lang een onderwerp van discussie in het onderzoek naar menselijke mobiliteit en reisgedrag. Big data zijn intuïtief beter, maar dit is niet altijd het geval. Big mobility data hebben betrekking op een groot aantal reizigers en reizen, maar er is weinig bekend over elke individuele reiziger en reis, en hun informatie soms moet worden samengevoegd omwille van privacy-overwegingen. Aan de andere kant bevatten gegevens uit reis-enquêtes, ondanks het feit dat ze slechts een kleine groep respondenten betreffen, over het algemeen een overvloed aan kenmerken over elke individuele reiziger, zoals leeftijd en attitudes, en elke reis, zoals het reisdoel. Als in een tabel met mobiliteitsdata elke rij één reiziger vertegenwoordigt en elke kolom één kenmerk, zouden big mobility data moeten worden beschreven als lang en dun, en "kleine" enquêtegegevens (Chen et al., 2016) als kort en breed.

Toch zijn big mobility data aantrekkelijk omdat ze kostenefficiënt, actueel en groot zijn in termen van steekproefgrootte. Daarom, gezien alle beperkingen en voordelen van big mobility data, wil dit proefschrift de volgende onderzoeksvraag beantwoorden:

In welke mate, en hoe, kunnen big mobility data het begrip van menselijk ruimtelijk gedrag bevorderen?

De belangrijkste bijdrage van dit proefschrift bestaat uit twee hoofdstrategieën die zijn gevolg om de onderzoeksvraag te beantwoorden. De eerste hoofdstrategie is om lange en dunne data breder te maken. Omdat het gebrek aan kenmerken het grootste obstakel is voor big mobility data om menselijk ruimtelijk gedrag te verklaren, zijn pogingen ondernomen om proxy variabelen te genereren voor reizigerssegmentatie en reiskarakterisering, hetzij uit big mobility data zelf (Hoofdstuk 2) of uit externe datasets (Hoofdstuk 3).

Gegeven een voorgeselecteerd activiteitendoel, is het bouwen van een locatiekeuzemodel met behulp van big mobility data vergelijkbaar met het bouwen van een locatiekeuzemodel met behulp van traditionele gegevens uit reis-enquêtes, met uitzondering van de afwezigheid van persoonlijke attributen. In hoofdstuk 2 doen we een eenvoudige aanname dat: de kenmerken van het gebied waar een persoon woont of werkt bepaalde kenmerken van deze persoon onthullen. We hebben uit longitudinale reisgegevens de woon- en werklocaties van elke reiziger gehaald en de woon-werkbalans van deze gebieden onthuild. Deze indicator dient dan als een vervangende variabele voor de persoonlijke kenmerken van elke reiziger en werd toegevoegd aan een discreet keuzemodel. Het bleek dat dit nieuwe model beter bij de data paste dan het model zonder persoonlijke kenmerken.

Hoofdstuk 3 was gebaseerd op dezelfde strategie, maar in plaats van alleen mobiliteitsgegevens te gebruiken om proxy variabelen te genereren, werden mobiliteitsgegevens gecombineerd met externe gegevensbronnen die een beeld gaven van de kenmerken van reizigers. Meer specifiek werd in de casestudy zowel gebruik gemaakt van traces van mobiele-telefoons om de bewegingen van reizigers te observeren, als ook van gegevens over mobiel internetgebruik (d.w.z. de frequentie van het bezoeken van elk type sites via smart phones) van dezelfde gebruikers. Daarnaast werd in de casestudy een clustermethode voorgesteld om soorten functionele stedelijke gebieden te bepalen met behulp van een andere externe gegevensbron, stedelijke point-of-interest gegevens. Uit de analyse blijkt dat het mobiele internetgebruik van reizigers statistisch gerelateerd is aan de soorten gebieden die zij het liefst bezoeken.

De tweede hoofdstrategie is geïnspireerd door de collaboratieve filteralgoritmen die gewoonlijk worden gebruikt om gebruikersvoorkeuren in aanbevelingssystemen te modelleren. Zonder gebruik te maken van specifieke proxy-variabelen zijn in hoofdstuk 4 en 5 data-gestuurde methoden geïmplementeerd, waarbij alleen wordt vertrouwd op empirische waarnemingen van een groot aantal mensen en waarbij geen op theorie gebaseerde aannames vooraf worden gedaan over het menselijk ruimtelijk gedrag. Deze benadering is intuïtief geschikt omdat historisch ruimtelijk gedrag zelf kan wijzen op heterogeniteit tussen individuen binnen een bepaalde groep reizigers en zo kan helpen voorspellingen te doen over hun toekomstige gedrag.

Hoofdstuk 4 implementeerde een buur-gebaseerd collaboratief filtering algoritme om de locatiekeuze voor flexibele activiteiten te voorspellen met behulp van metro smart card gegevens. Dit algoritme wordt veel gebruikt in aanbevelingssystemen om gebruikersvoorkeuren voor producten te voorspellen. Het zoekt naar zogenaamde “naaste buren” die vergelijkbare historische gedragspatronen delen met een individu en voorspelt dan de voorkeur van het individu op basis van het historische gedrag van de naaste buren. Onze casestudy toonde aan dat deze methode veelbelovend resultaten levert voor het voorspellen van reisgedrag. Omdat deze methode nog relatief onbekend is bij de meeste transportonderzoekers, is verder onderzocht kansrijk.

Als vervolg op de data-gedreven strategie, werd in Hoofdstuk 5 een matrix factorisatie model voorgesteld. Dit is een andere collaboratieve filtering methode die veel gebruikt

wordt in aanbevelingssystemen. Deze nieuwe methode kan vooral helpen bij het begrijpen van geaggregeerd ruimtelijk gedrag in termen van ruimtelijke interactie matrices. Het model is vergelijkbaar met een traditioneel op zwaartekracht gebaseerd model voor reisverdeling, maar flexibeler. In een casestudy laten we zien hoe het model herkomst-bestemmingsrittenmatrices in een metronetwerk kan voorspellen.

In het algemeen is er een wisselwerking tussen de strategieën die hetzij zijn gebaseerd op het toevoegen van proxy variablelen of vooral data gedreven zijn. Terwijl de data-gedreven strategieën profiteren van de flexibiliteit om gebruik te maken van patronen die in big data worden onthuld, zijn ze beperkt in hun interpreteerbaarheid en hun vermogen om de ontdekte patronen te generaliseren van geobserveerde reizigers of locaties naar niet-geobserveerde reizigers of locaties. Niettemin kan dit type methode voldoende nuttig zijn voor transportbedrijven om de dagelijkse gedragspatronen van hun bestaande klanten vast te leggen voor de dagelijkse operaties. De strategieën gebaseerd op het toevoegen van proxy variabelen zijn minder data-gestuurd, maar geschikter voor planning en beleidsvorming op langere termijn omdat het zich richt op het maken van voorspellingen in what-if scenario's met niet-waargenomen reizigers of locaties.

概述

人们每天都在通过空间从事各种活动。对活动类型和地点的选择被称为人类的空间行为。城市决策者需要了解土地使用和交通系统如何塑造人类的空间行为，以便设计更好的系统。过去，他们通过交通调查收集数据来了解人类的空间行为。今天，各种自动收集的大数据已经成为另一种可供选择的数据来源。

交通大数据与传统的交通调查数据孰优孰劣，一直以来是人类流动性和交通行为研究中具有争议性的话题。大数据在直觉上是更好的，但事实并非总是如此。交通大数据拥有大量出行样本，但对每个单独的出行者和出行细节知之甚少，更不用说有时出于隐私考虑，他们的信息必须被汇总。而传统的交通调查数据尽管只报告了一小部分受访者，但往往包括了关于每个出行者的丰富特征，如年龄和态度，以及每次出行的细节，如出行目的。假设数据中的每一行代表一个出行者或一次出行，每一列代表一个特征，交通大数据应该被更准确地描述为长而瘦的数据，而“小”调查数据(Chen et al., 2016) 则是短而宽。

尽管如此，交通大数据仍因其成本效益高、更新快、样本量大而受到青睐。因此，考虑到交通大数据的所有限制和优势，本论文旨在回答以下研究问题：

交通大数据在多大程度上能促进对人类空间行为的理解？如何促进？

本论文主要采取两个策略来回答这个研究问题。第一个主要策略是使长而瘦的数据变得更宽。由于缺乏特征是交通大数据解释人类空间行为的最大障碍，因此我们尝试利用交通大数据本身（第二章）或外部数据集（第三章）来产生用于描述出行者及其出行特征的替代变量。

给定一个活动目的，利用交通大数据建立地点选择模型与利用传统的交通调查数据建立地点选择模型没有太大区别，除了一点：交通大数据缺少个人属性。在第二章中，我们做了一个简单的假设：一个人居住或工作的地区的特征可以揭示这个人的特殊特征。我们从纵向出行数据中找到每个出行者的家庭和工作地点，并利用工作-住房平衡来揭示这些地区的一些特征。然后，这个指标作为每个出行者个人特征的代理变量，被添加到一个离散选择模型中。结果发现，这个新模型的表现优于没有任何个人特征的模型。

第三章是基于同样的策略，但选择了不同的路线。与其仅仅使用交通大数据本身来产生替代变量，不如将交通数据与描绘出行者特征的外部数据源相结合。具体来说，该案例研究不仅使用手机信令数据来观察出行者的行动，还使用了相同用户的移动互联网使用数据（即通过智能手机访问各类网站的频率）。此外，该案例研究提出了一种聚类方法，利用另一个外部数据源——城市兴趣点

数据确定城市功能区的类型。结果，我们发现出行者的移动互联网使用情况与他们喜欢去的地区类型有统计学上的关系。

第二个主要策略采取了一个相对更有突破性的方法，其灵感来自于推荐系统中常用于建模用户偏好的协同过滤算法。在不使用任何具体的代理变量的情况下，第四章和第五章实施了数据驱动的方法，只依赖于对人群历史行为的观察，不需要对人类空间行为的机制施加任何基于理论的先验假设。直观地说，这种方法可能会奏效，因为历史上的空间行为本身可以表明特定出行者群体的异质性，从而有助于对他们的未来行为做出预测。

第四章实现了一种最近邻的协同过滤算法，利用地铁交通卡数据预测活动的地点选择。这种算法已被广泛用于推荐系统中，以预测用户对产品的偏好。它寻找与个人有类似历史行为模式的所谓“邻居”，然后根据邻居的历史行为预测个人的偏好。我们的案例研究表明，这种方法在出行行为预测这个任务中表现相当好，建议进一步探索这种方法，因为大多数交通研究者对这种方法还比较陌生。

作为数据驱动策略的延续，第五章提出了矩阵分解模型，这是另一种在推荐系统中被广泛使用的协同过滤方法。这种新方法特别有助于理解空间OD 矩阵方面的总体空间行为。我们指出，这个模型几乎等同于传统的基于重力的交通分布模型，但更加灵活。在一个预测地铁网络的出发地-目的地OD 矩阵的案例研究中，我们证明了新模型在预测性能方面的优势。

总的来说，这两种策略之间需要权衡。虽然数据驱动的方法得益于它们在大数据中获取规律的灵活性，但它们的可解释性以及将发现的模式从观察到的出行者或地点推广到未观察到的出行者或地点的能力有限。尽管如此，这类方法对于交通运营公司来说是非常有用的，可以捕捉到他们现有用户的日常行为模式，用于日常运营。另一方面，关于添加代理变量的第一种主要策略不太受数据驱动，但仍然更适合于长期规划和政策制定，因为它们可以在涉及未观察到的出行者或地点的假设情况下进行预测。

About the author

Yihong Wang (汪一泓) was born on February 2, 1991 in Shanghai, China. He grew up in Shanghai, and completed his bachelor's study at the Shanghai Jiao Tong University in 2013.

Afterwards, Yihong came to the Netherlands to study his master's program in Transport and Planning at the Delft University of Technology. He finished the program and obtained his master's degree in June 2015. His master thesis was about using mobile phone traces to optimize road network in Senegal, and the work was awarded Transport Prize at the D4D Data Challenge organized by Orange Telecom and the NetMob conference.

In September 2015, Yihong started his PhD at the Delft University of Technology, funded by TRAIL Research School and NWO (Dutch Research Council). His main research topics (exactly the same as his research interests) can be described by the following key words: big data, machine learning, human mobility, travel behavior, and location choice. During the 4th year of his PhD, he participated in a side project where he studied the feasibility of remotely controlled (automated) vehicles for the manufacturing company Continental. In addition to his research, he also worked as a teaching assistant for several courses and a thesis supervisor.

Since September 2019, Yihong has been working as a data scientist at the online food delivery company Just Eat Takeaway.com in Amsterdam, the Netherlands, with a particular focus on solving geospatial data science problems. In 2020, he established the connection between the company and the researchers from Delft University of Technology and Erasmus University of Rotterdam, and helped them launch the NWO-funded project CUSTOMIZE: Customer-driven prescriptive analytics for logistics planning.

Journal papers

1. **Wang, Y.**, Correia, G.H.D.A., de Romph E. & Santos, B.F. (2020) Road network design in a developing dountry using mobile phone data: an application to Senegal. *IEEE Intelligent Transportation Systems Magazine*, 12, 2, 36-49.

2. **Wang, Y.**, Correia, G.H.D.A., van Arem, B., & Timmermans, H.J.P. (2018). Understanding travellers' preferences for different types of trip destination based on mobile internet usage data. *Transportation Research Part C: Emerging Technologies*, 90, 247-259.

3. **Wang, Y.**, Correia, G.H.D.A., de Romph, E., & Timmermans, H.J.P. (2017). Using metro smart card data to model location choice of after-work activities: an application to Shanghai. *Journal of Transport Geography*, 63, 40-47.

4. Wang, W., **Wang, Y.**, Correia, G.H.D.A., & Chen Y. (2020) A network-based model of passenger transfer flow between bus and metro: an application to the public transport system of Beijing. *Journal of Advanced Transportation*, 2020.

In-progress papers

1. **Wang, Y.**, Correia, G.H.D.A., van Arem, B., & Timmermans, H.J.P. Exploring a neighborhood-based collaborative filtering approach to modeling urban location preferences for flexible activities through metro smart card data. *Journal of Transport Geography*, under review.

2. **Wang, Y.**, Correia, G.H.D.A., van Arem, B., & Timmermans, H.J.P. A matrix factorization approach to modeling trip generators and spatial interactions. *Journal of Advanced Transportation*, under review.

3. Gao, K., Sun, L., Tu, H., Axhausen, K.W., **Wang, Y.** Inertia effects of past behavior in modal shift behavior: interactions, variations, and implications for demand estimation. *Transportation*, under review.

Book chapters

1. **Wang, Y.**, Correia, G.H.D.A., & van Arem, B. (2019). Relationships between mobile phone usage and activity-travel behavior: A review of the literature and an example. *Advances in Transport Policy and Planning*, 3, 81-105.

Conference presentations

1. **Wang, Y.**, Correia, G.H.D.A., van Arem, B., & Timmermans, H.J.P. (2019). A neighborhood-based collaborative filtering algorithm for secondary activity location choice prediction using smart card data. *Transportation Research Board 98th Annual Meeting (TRB)*, Washington, D.C., USA.

2. **Wang, Y.**, Correia, G.H.D.A., van Arem, B., & Timmermans, H.J.P. (2018). Understanding travellers' preferences for different types of trip destination based on mobile internet usage data. *International Conference on Travel Behavior Research (IATBR)*, Santa Babara, USA.

3. **Wang, Y.**, Correia, G.H.D.A., van Arem, B., & Timmermans, H.J.P. (2018). Using mobile internet usage behavior data to understand travel behavior. *TRAIL Congress*, *TRAIL Congress*, Utrecht, The Netherlands.

4. **Wang, Y.**, Correia, G.H.D.A., van Arem, B., Timmermans, H.J.P., & de Romph, E. (2017). Understanding muliday activity patterns based on mobile internet usage behavior. *NetMob*, Milan, Italy.

5. **Wang, Y.**, Correia, G.H.D.A., de Romph, E., & Timmermans, H.J.P. (2016). Using public transport smart card data to model location choice of after-work activity: an application to Shanghai. *TRAIL Congress*, Utrecht, The Netherlands.

6. **Wang, Y.**, Correia, G.H.D.A., de Romph, E. & Santos B.F. (2015). Use of mobile phone data for planning a road network: application to the country of Senegal. *EURO Working Group on Transportation (EWGT)*, Delft, The Netherlands.

7. **Wang, Y.**, Correia, G.H.D.A., de Romph, E. & Santos B.F. (2015). National and regional road network optimization for Senegal using mobile phone data *NetMob*, Cambridge, USA.

TRAIL Thesis Series

The following list contains the most recent dissertations in the TRAIL Thesis Series. For a complete overview of more than 275 titles see the TRAIL website: www.rsTRAIL.nl.

The TRAIL Thesis Series is a series of the Netherlands TRAIL Research School on transport, infrastructure and logistics.

Wang, Y., *Modeling Human Spatial Behavior through Big Mobility Data*, T2021/19, June 2021, TRAIL Thesis Series, the Netherlands

Coevering, P. van de, *The Interplay between Land Use, Travel Behaviour and Attitudes: a quest for causality*, T2021/18, June 2021, TRAIL Thesis Series, the Netherlands

Landman, R., *Operational Control Solutions for Traffic Management on a Network Level*, T2021/17, June 2021, TRAIL Thesis Series, the Netherlands

Zomer, L.-B., *Unravelling Urban Wayfinding: Studies on the development of spatial knowledge, activity patterns, and route dynamics of cyclists*, T2021/16, May 2021, TRAIL Thesis Series, the Netherlands

Núñez Velasco, J.P., *Should I Stop or Should I Cross? Interactions between vulnerable road users and automated vehicles*, T2021/15, May 2021, TRAIL Thesis Series, the Netherlands

Duivenvoorden, K., *Speed Up to Safe Interactions: The effects of intersection design and road users' behaviour on the interaction between cyclists and car drivers*, T2021/14, April 2021, TRAIL Thesis Series, the Netherlands

Nagalur Subraveti, H.H.S., *Lane-Specific Traffic Flow Control*, T2021/13, March 2021, TRAIL Thesis Series, the Netherlands

Beirigo, B.A., *Dynamic Fleet Management for Autonomous Vehicles: Learning- and optimization-based strategies*, T2021/12, March 2021, TRAIL Thesis Series, the Netherlands

Zhang, B., *Taking Back the Wheel: Transition of control from automated cars and trucks to manual driving*, T2021/11, February 2021, TRAIL Thesis Series, the Netherlands

Boelhouwer, A., *Exploring, Developing and Evaluating In-Car HMI to Support Appropriate use of Automated Cars*, T2021/10, January 2021, TRAIL Thesis Series, the Netherlands

Li, X., *Development of an Integrity Analytical Model to Predict the Wet Collapse Pressure of Flexible Risers*, T2021/9, February 2021, TRAIL Thesis Series, the Netherlands

Li, Z., *Surface Crack Growth in Metallic Pipes Reinforced with Composite Repair System*, T2021/8, January 2021, TRAIL Thesis Series, the Netherlands

Gavriilidou, A., *Cyclists in Motion: From data collection to behavioural models*, T2021/7, February 2021, TRAIL Thesis Series, the Netherlands

Methorst, R., *Exploring the Pedestrians Realm: An overview of insights needed for developing a generative system approach to walkability*, T2021/6, February 2021, TRAIL Thesis Series, the Netherlands

Walker, F., *To Trust or Not to Trust? Assessment and calibration of driver trust in automated vehicles*, T2021/5, February 2021, TRAIL Thesis Series, the Netherlands

Schneider, F., *Spatial Activity-travel Patterns of Cyclists*, T2021/4, February 2021, TRAIL Thesis Series, the Netherlands

Madadi, B., *Design and Optimization of Road Networks for Automated Vehicles*, T2021/3, January 2021, TRAIL Thesis Series, the Netherlands

Krabbenborg, L.D.M., *Tradable Credits for Congestion Management: support/reject?*, T2021/2, January 2021, TRAIL Thesis Series, the Netherlands

Castelein, B., *Accommodating Cold Logistics Chains in Seaport Clusters: The development of the reefer container market and its implications for logistics and policy*, T2021/1, January 2021, TRAIL Thesis Series, the Netherlands

Huang, B., *The Influence of Positive Interventions on Cycling*, T2020/20, December 2020, TRAIL Thesis Series, the Netherlands

Xiao, L., *Cooperative Adaptive Cruise Control Vehicles on Highways: Modelling and Traffic Flow Characteristics*, T2020/19, December 2020, TRAIL Thesis Series, the Netherlands

Polinder, G.J., *New Models and Applications for Railway Timetabling*, T2020/18, December 2020, TRAIL Thesis Series, the Netherlands

Scharpff, J.C.D., *Collective Decision Making trough Self-regulation*, T2020/17, November 2020, TRAIL Thesis Series, the Netherlands

Guo, W. *Optimization of Synchromodal Matching Platforms under Uncertainties*, T2020/16, November 2020, TRAIL Thesis Series, the Netherlands

Narayan, J., *Design and Analysis of On-Demand Mobility Systems*, T2020/15, October 2020, TRAIL Thesis Series, the Netherlands

Gong, X., *Using Social Media to Characterise Crowds in City Events for Crowd Management*, T2020/14, September 2020, TRAIL Thesis Series, the Netherlands

Rijal, A., *Managing External Temporal Constraints in Manual Warehouses*, T2020/13, September 2020, TRAIL Thesis Series, the Netherlands

Alonso González, M.J., *Demand for Urban Pooled On-Demand Services: Attitudes, preferences and usage*, T2020/12, July 2020, TRAIL Thesis Series, the Netherlands

Alwosheel, A.S.A., *Trustworthy and Explainable Artificial Neural Networks for choice Behaviour Analysis*, T2020/11, July 2020, TRAIL Thesis Series, the Netherlands