Data-driven Analysis and Modeling of Passenger Flows and Service Networks for Public Transport Systems

Luo, Ding

DOI
10.4233/uuid:ebc2f602-a65b-491c-805f-9fb4f37cb104

Publication date
2020

Document Version
Final published version

Citation (APA)
Luo, D. (2020). *Data-driven Analysis and Modeling of Passenger Flows and Service Networks for Public Transport Systems*. [Dissertation (TU Delft), Delft University of Technology]. TRAIL Research School. https://doi.org/10.4233/uuid:ebc2f602-a65b-491c-805f-9fb4f37cb104

TRAIL

TRAIL THESIS SERIES T2020/2

Summary

Public transport plays an increasingly important role in solving mobility challenges. Despite the considerable amount of data currently being generated and collected for public transport systems, our capability of using these data for improving planning and operations is still limited. To this end, this thesis is dedicated to developing methods and models for translating high-volume data from various sources into novel knowledge and insights that can be used to improve public transport planning and operations.

About the Author

Ding Luo conducted his PhD with the department of Transport and Planning at Delft University of Technology from March 2016 to February 2020. He received his bachelor's degree and master's degree from Beijing Jiaotong University and KTH Royal Institute of Technology, respectively.

TRAIL Research School ISBN 978-90-5584-258-2

Erasmus University Rotterdam · Radboud University · rijksuniversiteit groningen · TU Delft Delft University of Technology · University of Twente. · TU/e Technische Universiteit Eindhoven University of Technology

Ding Luo — Data-driven Analysis and Modeling of Passenger Flows and Service Networks for Public Transport Systems

Data-driven Analysis and Modeling of Passenger Flows and Service Networks for Public Transport Systems

Ding Luo

Data-driven Analysis and Modeling of Passenger Flows and Service Networks for Public Transport Systems

Ding Luo

This doctoral project received financial support from the SETA project
funded by the European Union's Horizon 2020 research and innovation program.

Data-driven Analysis and Modeling of Passenger Flows and Service Networks for Public Transport Systems

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
Friday 21 February 2020 at 12:30 o'clock
by

Ding LUO

Master of Science in Built Environment,
KTH Royal Institute of Technology, Sweden
born in Tianshui, Gansu Province, China

This dissertation has been approved by the promoters.

Composition of the doctoral committee:

Rector Magnificus	Chairperson
Prof.dr.ir. J.W.C. van Lint	Delft University of Technology, promoter
Dr. O. Cats	Delft University of Technology, promoter

Independent members:

Prof.dr.ir. R.E. Kooij	Delft University of Technology, the Netherlands
Prof.dr. M. Trépanier	Polytechnique Montréal, Canada
Prof.dr. M. Munizaga	University of Chile, Chile
Dr. E. Jenelius	KTH Royal Institute of Technology, Sweden
Dr.ir. N. van Oort	Delft University of Technology, the Netherlands

Reserve member:

Prof.dr.ir. S.P. Hoogendoorn	Delft University of Technology, the Netherlands

TRAIL Thesis Series no. T2020/2, the Netherlands Research School TRAIL

TRAIL
P.O. Box 5017
2600 GA Delft
The Netherlands
E-mail: info@rsTRAIL.nl

ISBN: 978-90-5584-258-2

Printed in the Netherlands

Do not go gentle into that good night
Old age should burn and rave at close of day
Rage, rage against the dying of the light

Though wise men at their end know dark is right
Because their words had forked no lightning they
Do not go gentle into that good night

- Dylan Thomas

Acknowledgements

PhD students at TU Delft are normally expected to obtain their doctorates – which means to also complete defending their these publicly – within 4 years. That is why we are offered a four-year contract with the university. However, not even 5% of the PhD candidates at TU Delft can realize this according to the latest survey[1]. Big challenge!

Nevertheless, I have somehow managed to become a member of this niche community at the end of the day (3 years, 11 months, 21 days). What an achievement! But wait! Don't get me wrong and consider this as self-boasting! Frankly, my motivation for saying so is to emphasize the tremendous supports and supervision I have received from my supervisors, Hans and Oded, during this process. Given my own experience, I firmly believe that the key to a successful PhD tremendously depends on the synergy between a PhD's personal effort and his/her supervisors' inputs from different dimensions. That being said, I want to first and foremost express my deepest gratitude to Hans and Oded for their extensive guidance, supports, encouragement and patience in the past years. I have enjoyed working with you a lot, and I do hope that you have felt the same way and never regretted bringing me to Delft:)

I want to further extend my gratitude to Prof. Graham Currie and Dr. Niels van Oort. Graham, many thanks for hosting me in Melbourne and supporting me on the Monash SASXP (I still remember your joke on this bizarre abbreviation) application. Niels, I truly appreciate all the chats and discussions that we have had in various places, especially that late-night Uber ride in the middle of nowhere in Melbourne. You have been a great mentor to me and I am very honored that you are in the evaluation committee of this dissertation.

Deep gratitude goes to the rest committee members of this dissertation: Rob, Marcela, Martin, Erik and Serge. The fact that I actually have met and talked to most of you in person at academic conferences makes me feel much honored to have you assess this dissertation.

Next, I want to acknowledge my dear Dittlab fellows (and some of their partners), Panchamy & Jerry, Tín & Huong, Léonie & Justin, Ehab & Shahad, Zahra, Sanmay, Kristel, Guopeng, Simeon, Peter. This has truly been one of the birthplaces of joy

[1] https://www.delta.tudelft.nl/article/why-phds-are-not-obtaining-their-doctorates-time

(and pain) of my (our) life. Of course, I will never forget the visitors who contributed to making these joy and pain too, Clélia & Etienne, Loïc, Rafael, Juan, Alan, Kota, and Julian. Very distinct acknowledgments must go to Panchamy, whose last name, Krishnakumari, could never be correctly pronounced by me. We have been sitting next to each other for four years at TU Delft, but it is unbelievable that I first received your help at KTH back in 2014. It was amazing that we ended up doing our PhDs together in Delft. In this sense, the SETA project was indeed a great success.

My gratitude is also extended to a long list of lovely people from the department of Transport and Planning. I have benefited significantly from this group in terms of both academic and social activities. I appreciate all the group lunches, after-lunch coffee tours and spontaneous get-togethers. To my "indigenous" colleagues:Danique, Paul, Tim, Pablo, Lara, Marie-Jette, Martijn, Arjan, het is jammer dat mijn Nederlands nog niet goed genoeg voor meer interessant gesprekken is. Ik dank jullie voor het oefenen met jullie. Bedankt voor al het advies en het boek, Danique!

Many thanks for other colleagues: Florian, Giulia, Alexandra, Alphonse, Nikola, Joelle, Niharika, Marko, Menno, Bahman, Konstanze, Malvika, Peyman, Rafael, Nejc, Jishnu. Special thanks go to my amiga María and amigo Xavi for somehow building some true friendship between a Chinese (cabrón) and Spanish people. Gracias, María, for all the advice, company, cares and sharing. The fact that we have traveled together to the US, Australia, China is simply amazing. And I believe the end of our PhDs will not be the end of more adventures ahead. Xavi, many thanks for allowing me to share my troubles with you and cheering me up. Catching up with you has always been a nice thing to do.

I would like to acknowledge the T&P Chinese group for their support and help, Vincent, Yihong, Yufei, Han, Xiao & Qu, Yongqiu & Hongrui, Meng, Lin & Zhen, Yaqing & Lan, Meiqi, Pengling. Special thanks go to Vincent and Yihong for countless conversations and activities that are worthwhile memorizing. I also want to extend my gratitude to some friends who have provided me with company and help along this journey around the world, He, Mengxue, Wenxin, Lin, Yi, Boyao, Rong, Jianrong, Wenhua, Yuqing.

Last but not least, my ultimate debt of gratitude goes to my family back in Tianshui where I was born and raised. 感谢父母赠予我生命，感谢奶奶与姑姑也同时陪伴我长大成人，自离开家以来的经历让我逐渐理解亲情确是这世间最深的缘分。生命的延续需要共识，但异见与争执也不可或缺。感谢你们无限的付出与支持，愿我们能分享更多生命的美好。

Ding
Delft, December 2019

Contents

List of Figures

List of Tables

List of Acronyms and Abbreviations

AFC	Automatic fare collection
AVL	Automatic vehicle location
CDF	Cumulative distribution function
GIS	Geographical information systems
GTC	Generalized travel cost
GTFS	General transit feed specifications
OD	Origin-destination
PC	Principal component
PCA	Principal component analysis
PT	Public transport
PTN	Public transport network
TAZ	Traffic analysis zone

Chapter 1

Introduction

Public transport plays an increasingly important role in solving mobility challenges, especially in densely populated metropolitan areas. Despite the considerable amount of data currently being generated and collected for public transport systems, our capability of using these data for improving planning and operations is still limited. To this end, this thesis presents research that has been dedicated to developing methods and models for translating high-volume data from a variety of sources into novel knowledge and insights that can be used to improve public transport planning and operations. Our research makes multiple scientific contributions to the field of transport and mobility, and is of high societal relevance in terms of developing more advanced mobility systems.

In this opening chapter, we first discuss the research motivation in section 1.1, and then introduce the overarching research objective along with specific research questions in section 1.2. Section 1.3 briefly describes the research context of our research. We further elaborate on our scientific contributions in section 1.4, followed by the discussion on the societal relevance of our research in section 1.5. This chapter is finalized with the outline of the thesis in section 1.6.

1.1 Research motivation

Many places worldwide, especially those densely populated cities or metropolises, are currently facing a critical challenge: how can we develop mobility systems that can meet people's increasing needs for traveling while maintaining high sustainability, reliably and cost efficiency? Despite great expectations on some technologies that might fundamentally change our mobility systems, e.g., autonomous vehicles, it is still believed that public transport (PT) will keep its dominant role during and after this mobility revolution (Currie, 2018). Building more advanced PT systems, therefore, is arguably one of the most tangible solutions to the mobility challenge our society is dealing with.

One of the major obstacles that hinder further development of PT systems, however, is the lack of insight into the complex dynamics of passengers and vehicles. It is thus difficult to improve the performance of current PT systems as well as travelers' experience without advanced planning and operations. To address this challenge, more research is needed for broadening and deepening existing knowledge of PT systems, which is currently still lagging behind the research on car traffic.

Fortunately, the growing availability of high-volume data from a variety of sources is offering an unprecedented opportunity for resolving this issue. These new data may help to significantly alter the so-called "assumption-rich and data-poor" (Vlahogianni et al., 2015) situation that has lasted for decades in the PT research community. It unlocks the potential for understanding PT dynamics regarding demand and supply based on various measurements. Nonetheless, data do not automatically turn into valuable insights and knowledge by themselves, which calls for new models and methods that can translate them into the desired knowledge. This thesis is therefore dedicated to bridging this research gap.

1.2 Research objective and questions

The overarching research objective of this thesis is to **develop methods and models for translating high-volume data from a variety of sources into novel knowledge and insights that can be used to improve public transport planning and operations**.

Centering around this research objective, we further formulate the following research questions (RQs) that have not been addressed by existing studies.

- ***RQ1: How can we generate more information-rich profiles of PT vehicles containing positions and onboard occupancy based on prevalent PT data sources?*** (**Chapter 3**)

- ***RQ2: How can we reduce the high dimensionality of passenger flows for large-scale analysis and modeling?*** (**Chapter 4**)
- ***RQ3: How can we construct zone-to-zone OD matrices for PT systems using data-driven techniques?*** (**Chapter 5**)
- ***RQ4: How can we analyze the accessibility of public transport networks in a more transferable and efficient way?*** (**Chapter 6**)
- ***RQ5: Can passenger flows be estimated solely based on the network properties of PT systems?*** (**Chapter 7**)

1.3 Research context

This thesis was supported by SETA[1], a research and innovation project funded by the European Union's Horizon 2020 program. The project was performed from February 2016 to January 2019 with 14 partners from the United Kingdom, Spain, Poland, the Netherlands and Italy. SETA envisages developing a ubiquitous data and service ecosystem for better metropolitan mobility. The overall goal of the project is described as follows.

SETA creates technologies and methodologies set to change the way mobility is organised, monitored and planned in large metropolitan areas. The solutions are based on large, complex dynamic data from millions of citizens, thousands of connected cars, thousands of city sensors and hundreds of distributed databases.

1.4 Scientific contributions

Overall, this thesis makes scientific contributions related to data-driven methods and models for better understanding passenger flows and service networks in PT systems. Specific contributions of each chapter are further detailed as follows.

- **Developing a method for constructing PT vehicles' trajectories with on-board occupancy with multiple data sources (Chapter 3)**

  The contribution here is twofold. First, the research systematically identifies the issues related to each and the combination of different data sources, namely AFC, AVL, and GTFS, for performing this task. Although the study is performed with the data from the Netherlands, the identified issues are universal and are highly beneficial for researchers and practitioners encountering different data formats yet with similar difficulties. Second, a method for solving these issues in

[1] https://cordis.europa.eu/project/rcn/199852/factsheet/en

a sequential manner is developed. The complexity of approaches and algorithms in each step can vary depending on the availability of information.

This contribution has led to the following journal article:

Luo, D., Bonnetain, L., Cats, O. & van Lint, H. (2018) Constructing spatiotemporal load profiles of transit vehicles with multiple data sources. *Transportation Research Record*, 2672(8), 175-186.

- **Developing a method based on principal component analysis (PCA) for understanding network-wide PT passenger flows (Chapter 4)**

  This contribution pertains to developing multivariate analytical techniques on PT passenger flows. It also shows the potential of incorporating PCA into applications such as flow anomaly detection and short-term forecasting.

  This contribution has led to the following conference paper:

  Luo, D., Cats, O. & van Lint, H. (2017) Analysis of network-wide transit passenger flows based on principal component analysis. In *Proceedings of the 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 744-749.

- **Developing a method based on the *k*-means algorithm for clustering PT stops using additional information of passenger flows (Chapter 5)**

  The proposed method provides a new data-driven perspective for zoning PT stops considering both passenger flow and spatial distance patterns. The method allows for obtaining clusters that are, on the one hand, sufficiently large to enable the consideration and modeling of travel alternatives between parts of the network; and, on the other hand, are compact enough to include only viable alternatives and support fine-grained demand estimation.

  This contribution has led to the following journal article:

  Luo, D., Cats, O. & van Lint, H. (2017) Constructing transit origin-destination matrices with spatial clustering. *Transportation Research Record*, 2652(1), 39-49.

- **Developing a method for integrating network science and PT accessibility analysis for comparative assessment (Chapter 6)**

  This contribution is twofold. First, the study proposes a new method based on network science for computing PT accessibility measured as the average travel impedance. Second, the study uses eight tram networks worldwide in the case study to demonstrate the proposed method. The comparative assessment reveals new findings about the accessibility of these tram networks, particularly with insights into how different travel components (e.g., in-vehicle travel times and waiting and transfer times) specifically contribute to the variance in accessibility across different networks. Such latitudinal comparative assessments can provide

additional insights into the public transport network design, benchmark and planning, but are still scarce in the current literature due to the requirements imposed by existing methods that heavily rely on geographical information systems.

This contribution has led to the following journal article:

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

- **Conducting a pioneering investigation into the relation between passenger flow distribution and network properties in PT systems (Chapter 7)**

  We conducted this empirical investigation based on the observed data from the tram networks of two Dutch cities, namely The Hague and Amsterdam. We conclude that the selected network properties can indeed be used to approximate passenger flow distribution in public transport systems to a reasonable extent. The significance and relevance of this study stems from two aspects: (1) our finding provides a parsimonious alternative to existing passenger assignment models that require many assumptions on the basis of limited data; (2) the resulting model offers efficient quick-scan decision support capabilities that can help transport planners in tactical planning decisions.

  This contribution has led to the following journal article:

  Luo, D., Cats, O. & van Lint, H. (2019) Can passenger flow distribution be estimated solely based on network properties in public transport systems? *Transportation*. https://doi.org/10.1007/s11116-019-09990-w.

1.5 Societal relevance

Developing efficient, reliable and sustainable mobility systems is an important challenge for many societies worldwide. Over the past yeas, the greatly increased availability of PT data has opened an unprecedented opportunity for operators to achieve more advanced and informed planning and operations, but many of them are still struggling to do so due to the lack of knowledge about how data from a variety of sources can be integrated for more insightful analyses and modeling. This thesis therefore helps to unlock the potential of data for improving PT planning and operations, ultimately contributing to addressing societal mobility challenges with more advanced PT systems. More specifically, this thesis offers the following scientific outcomes that can benefit the PT industry.

- Techniques and algorithms for converting multi-source PT data to crucial information needed for planning and operations.
- An efficient approach which allows for comparative assessment of PT accessibility based on scheduled services.

- An efficient approach for zoning networks based on both spatial and passenger flow patterns.
- New perspectives for estimating passenger flows and evaluating the performance of PT systems combining both demand and supply.

1.6 Thesis outline

The remainder of this thesis comprises seven chapters, with the structure displayed in Figure 1.1. In addition to all the individual chapters with their specific topics, a layer consisting of three themes (colorful boxes with dashed lines) is inserted in the background to illustrate the position of each individual chapter and the connections among them. Note that some chapters (e.g., **Chapter 6** and **Chapter 7**) intersect more than one theme. In what follows, the chapters will be introduced along with the themes.

Chapter 2 presents fundamental concepts, data and frameworks for the PT research of this thesis. We introduce PT from a system perspective, characterizing it with a couple of essential components. We also discuss current PT data from multiple sources, demonstrating them with an example of The Hague, the Netherlands. The frameworks for dealing with data and performing data-driven research are then presented.

Chapter 3 describes how onboard occupancy for PT vehicles can be inferred using three different data sources including AFC, AVL and GTFS. It demonstrates the issues and corresponding solutions when converting raw multi-source data into critical information needed for further PT analysis and modeling. This chapter forms a crucial part of the **Information Generation** theme, which provides direct input for **Chapter 7**.

The theme of **Passenger Flows** pertaining to the demand side of PT systems underlies both **Chapter 4** and **Chapter 5**. In fact, both chapters attempt to address the same challenge – i.e., high-dimensional passenger flow data over time and space – when performing data-driven analysis and modeling of passenger flows at the entire network level. **Chapter 4** resolves this challenge by applying a statistical technique (i.e., PCA). The ultimate goal is to effectively retain the variance contained in the data when projecting them from the original high-dimensional space to a low-dimensional one. **Chapter 5** then endeavors to solve this problem more effectively in the context of PT planning. By clustering PT stops based on both spatial and demand patterns, the issue of high dimensionality for analysis and modeling is largely alleviated because the number of clusters is much smaller than the number of PT stops. Moreover, it often makes sense to investigate PT demand on a zonal basis, given travelers' preference in choosing different stops and routes for their journeys.

Chapter 6 contributes to both the **Information Generation** and the **Service Networks** themes. It does so by integrating network science with PT accessibility analysis. Regarding the contribution to the former theme, this chapter depicts how PTNs can be

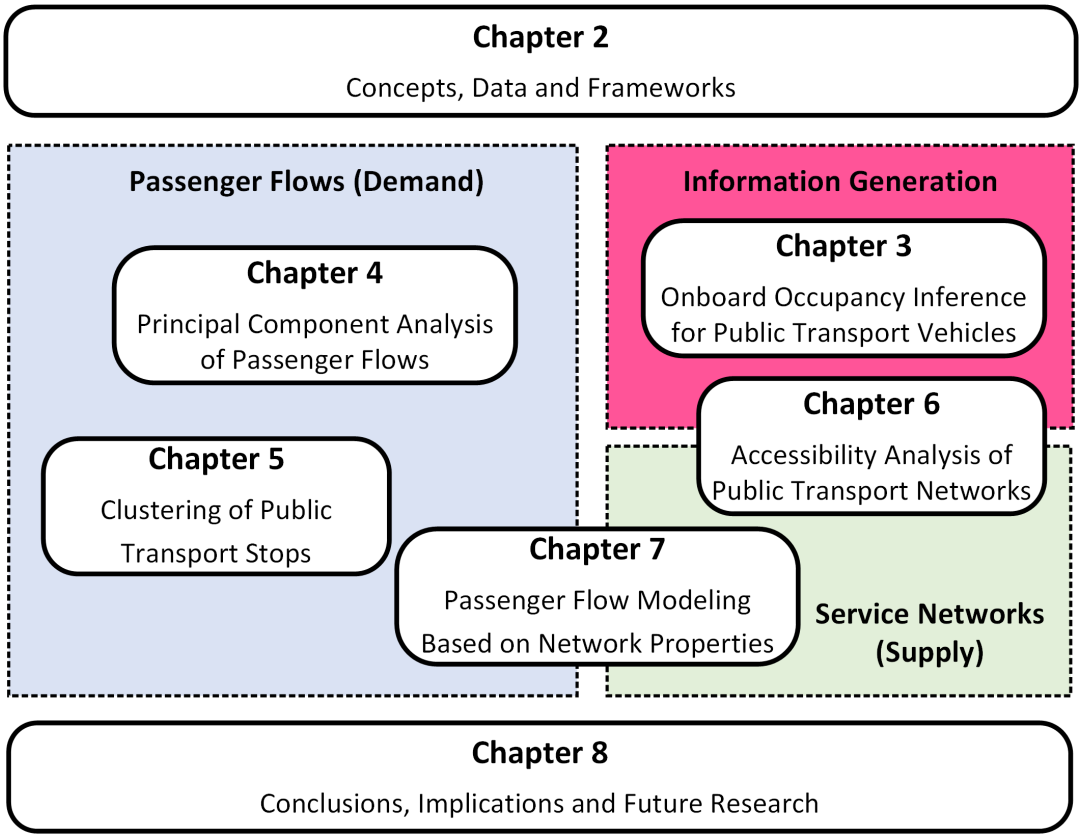


Figure 1.1: Structure of the thesis.

concisely represented as graphs, with meaningful weights derived from schedules contained in GTFS data. Furthermore, applying network science to PTN studies provides an effective way to develop insights into PT service networks (i.e., supply), thus contributing to the latter theme.

Chapter 7 presents an investigation concerning both themes. In particular, this chapter is devoted to determining whether passenger flow distribution can solely be estimated based on network properties of PT systems. It thus examines the relation between passenger flows and service networks.

In **Chapter 8**, we draw overall conclusions and further discuss the implications for practice of our research. We finalize this dissertation by giving recommendations for future research.

Chapter 2

Concepts, Data and Frameworks

Before presenting specific methodological advancements, we first introduce important concepts, data and frameworks that lay the foundation of current PT research in this chapter. We start by characterizing PT from a system perspective in section 2.1, in which the essential components and terminology are specified. Next we describe current multi-source PT data in section 2.2 with a concrete example of The Hague, the Netherlands because the data collected from The Hague's urban PT system have been extensively used throughout the research presented in this thesis. We introduce a framework for processing and fusing these multi-source PT data in section 2.3, followed by the illustration of another framework for data-driven PT research that is designed for enhancing planning and operations in section 2.4.

2.1 Public transport from a system perspective

We view PT from a system perspective throughout the research in this thesis, which is illustrated in Figure 2.1 along with its essential components. On the demand side, *(passenger) journeys* constitute the first component, which are specified by origin and destination stops without considering the access and egress to PT systems. A journey can contain either one or multiple *(passenger) rides*. When there are multiple rides, it is necessary for passengers to perform *transfers*. On the supply side, there are networks (i.e., *routes*) and services (i.e., *[vehicle] trips*). Each trip is executed by a PT vehicle, and the set of positions of this vehicle over time is referred to as its *trajectory*. There are scheduled and actual trips and trajectories, which will be introduced in the following sections. A summary of these components and their definitions is provided below, and these terms will be used throughout this thesis.

- **(Passenger) Rides**: The movement of a passenger using a **single** PT vehicle, i.e., bus, tram or metro. The ride begins at the stop where the passenger boards the vehicle and ends at the stop where the passenger alights from the vehicle. A single ride contains **no transfers**.
- **(Passenger) Journeys**: The movement of a passenger from an **origin** to a **destination**. In this context, the origin is assumed to be the **first PT stop** at which the passenger enters the network. Likewise, the destination location is assumed to be the **last PT stop** at which the passenger exits the network. There can be several rides included in one journey with transfer activities connecting them.
- **(Passenger) Transfers**: The movement of a passenger from one PT vehicle to another. There are different ways of performing a transfer. For example, a passenger can transfer at the same stop to a different route, or he/she can do so by walking to another nearby stop.
- **Routes**: A route is a pre-defined **sequence of stops**, and hence directed. It is also interchangeably referred to as **"line"** in both research and practice.
- **(Vehicle) Trips**: A trip is the movement of a PT vehicle through (part of) a route, i.e., a pre-defined sequence of stops. It is interchangeably referred to as **"run"** in both research and practice.
- **(Vehicle) Trajectories**: PT vehicles' **positions** over time and space. From the trajectories, we should be able to obtain PT vehicles' actual departure and/or arrival times at individual stops for all the trips.

These components are essential for any PT system. The following section will further describe how current multi-source PT data can be used to quantitatively characterize these components by performing processing, fusion and inference techniques.

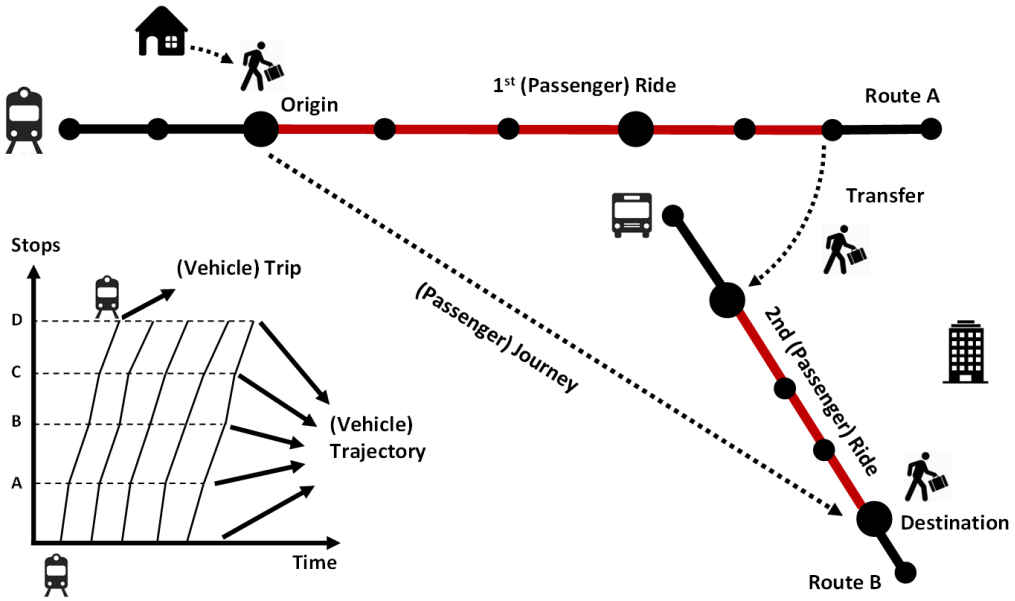


Figure 2.1: Schematic illustration of a PT system with essential components.

2.2 Public transport data

2.2.1 Current multi-source data

The availability of PT data has dramatically increased over the past decades owing to technological advancements. Existing reviews and taxonomy on PT data can be found in recent literature, such as Koutsopoulos et al. (2017) and Welch & Widita (2019). In this thesis, we choose to only introduce the following five different types of PT data which can be used in the proposed data processing and fusion framework.

The first primary data source pertains to **automatic fare collection (AFC)** data, which are also sometimes referred to as smart card data because these systems are now commonly dependent on contactless smart cards to record entry and/or exit information (geolocations and time stamps) of individual passengers. This data source significantly improves the capability of observing passenger demand patterns to a sufficiently accurate extent. For a detailed introduction of AFC data, readers are referred to the review paper by Pelletier et al. (2011). We extensively use this data source throughout the research in this thesis.

Automatic vehicle location (AVL) data are the second most important source for PT research. They provide detailed information about vehicles' actual trajectories. AVL data have been widely applied in various studies, and a literature survey on this topic has been presented by Moreira-Matias et al. (2015). Our research has also greatly benefited from AVL data.

The third data source is **general transit feed specification (GTFS)**, which is a standard data format initiated by Google (2019). It allows for systematically storing and

sharing PT schedules and associated geographic information through a series of text files. The significance of this data source has been increasingly recognized by both researchers and practitioners over the past years. This data source also plays an important role in our research.

The last two types of data are **automatic passenger counting (APC)** and **mobile crowdsensing**, which are not used in the research presented in this thesis but are mentioned because of their potential complementary roles in the overall data processing and fusion framework. APC data were commonly used as the major source for estimating demand (e.g., Ji et al., 2014, 2015), but have received much less attention since AFC data prevailed. Mobile crowdsensing data, on the contrary, are very promising given rapid growth in the mobile phone technology.

2.2.2 Dutch public transport data: an example of The Hague

We now introduce the Dutch PT data based on the example of The Hague, the third-largest city in the Netherlands, because their data have been extensively leveraged throughout the research presented in this thesis. In particular, the management and quality of the Dutch PT data are also leading worldwide, making the Netherlands one of the most suitable places to perform data-driven PT research and further to test data-driven enhancement of PT planning and operations. Operated by the local company HTM, its urban PT system consists of 12 tram and 8 bus routes serving more than 900 stops. Figure 2.2 provides a sketch of the PT system of The Hague.

AFC data

The Dutch AFC data are collected based on a nationwide smart card system, which is called *OV-chipkaart* in Dutch. A brief introduction of the history of this system has been presented by van Oort et al. (2015a). An important feature of the Dutch smart card system is that passengers are required to check in and check out for every single ride of a journey for fare calculation, except when transferring within the national railway system and metro systems. Missing check-out will cause a much higher fare, hence providing a strong incentive for passengers to check out. Consequently, records of travelers' origin and destination are fairly complete and accurate without the need to infer passenger alighting stops. However, individual rides still need to be combined to obtain an OD matrix, which is an important input for a variety of offline applications.

In this thesis, a one-month AFC data set for The Hague's PT system has been primarily used. The data set covers the whole month of March 2015 and contains close to 8 million validated records. Each record characterizes a single ride with anonymous card ID, route ID, date, stop ID and times of check-in and check-out. The information of trip ID and vehicle ID, which would allow for the connection to the AVL data, is however not available in this data set.

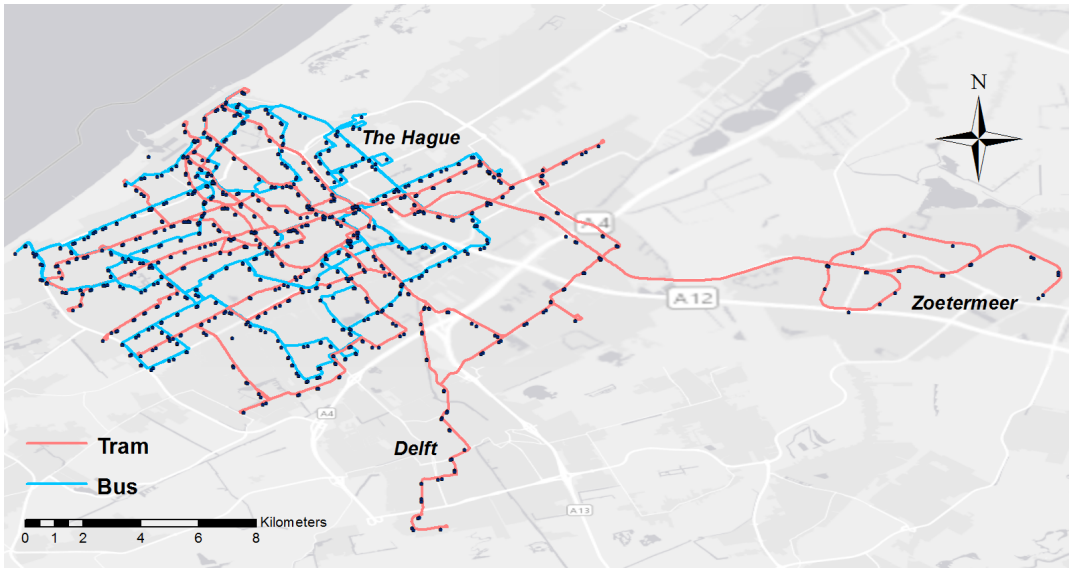


Figure 2.2: Illustration of the PT system of The Hague operated by HTM.

AVL data

The Dutch AVL data are well stored and managed at a national level as well. Details can be further found in the study by van Oort et al. (2015b). Our research used a one-month AVL data set that both spatially and temporally aligns with the AFC data set described above. It contains over 22 million records in an event-based format. Individual trips are distinguished by a unique trip ID within an operating day. Besides route ID, stop ID and vehicle ID, each row of data is characterized as an event (e.g., start of a trip, on-route, arrival, dwell, departure, and end of a trip) with the corresponding timestamp. In addition, deviation from the scheduled time is indicated in the data under "punctuality".

GTFS data

The Dutch GTFS data are also well managed at the national level. They are updated on a daily basis[1] by a non-profit organization called Stichting OpenGeo[2]. The feeds are created from the open data files published by local PT operators under an open license. The website freely provides up-to-date static GTFS data for the whole country, with the availability of GTFS-Realtime updates for some areas.

2.3 Framework for processing and fusing multi-source data

Based on the multi-source PT data that have been briefly introduced above, we propose a framework for processing and fusing these data, which can serve as a guideline for

[1] http://gtfs.ovapi.nl/
[2] https://ndovloket.nl/

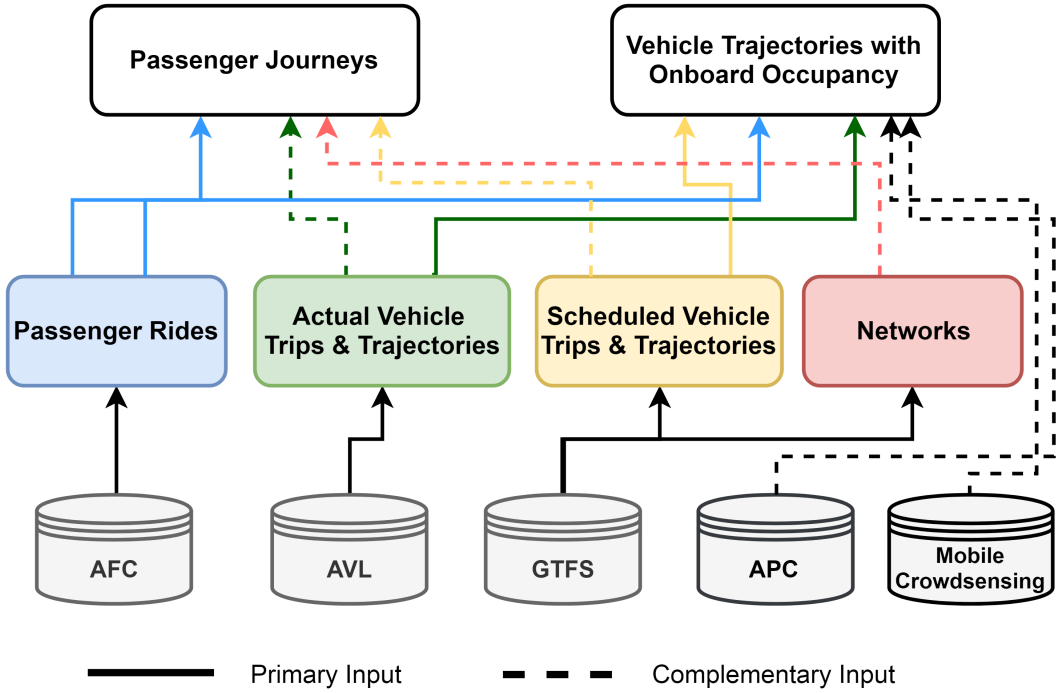


Figure 2.3: Illustration of the proposed framework for processing and fusing multi-source PT data.

obtaining key information for PT analysis and modeling. As Figure 2.3 shows, this framework consists of three layers, with the bottom, middle, and top ones representing the original data sources, the information obtained by processing standalone data sources, and the information obtained by fusing multiple data sources, respectively. In the diagram, primary inputs for derived information are marked by solid lines with arrows, whereas complementary inputs are marked by dashed lines.

There are four components in the middle layer, including passenger rides, actual vehicle trips and trajectories, scheduled vehicle trips and trajectories, and networks. The diagram indicates from which original data source each component can be obtained. However, it should be noted that the effort for obtaining these components varies depending on the quality and information richness of the original data sources. This is particularly the case for obtaining passenger rides based on AFC data and obtaining actual vehicle trips and trajectories based on AVL data. For example, passengers' alighting stops have to be inferred if check-out is not required by the AFC system due to the flat-fare scheme. A number of studies have been dedicated to developing the needed inference algorithm, such as Gordon et al. (2013); Alsger et al. (2016); Yan et al. (2019). If the fare collection is based on distance, then normally this inference step can be omitted because both check-in and check-out records would be stored, which makes it easy to generate passenger rides. For AVL data, matching and filtering techniques become necessary for deriving stop-based arrival and departure times if the system logs raw global positioning system (GPS) coordinates over time only (Cathey & Dailey, 2003).

Obtaining the components on the top layer, including passenger journeys and vehicle trajectories with onboard occupancy, requires fusing multiple outputs from the middle layer. For the former, it is necessary to develop chaining algorithms to link passengers' rides (e.g., Gordon et al., 2013; Alsger et al., 2016; Kumar et al., 2018). The key is to determine transfer activities based on spatial and temporal constraints. For the latter, the critical effort pertains to integrating multiple data sources through common indices. However, when the common indices are unavailable, inference methods will need to be developed to match different sources.

2.4 Framework for data-driven public transport research

We further present a framework for data-driven PT research in Figure 2.4, of which ultimate goal is to contribute to the improvement of PT planning and operations. Note that our proposed framework is a modified edition based on the ones presented by Koutsopoulos et al. (2017, 2019). As the diagram shows, a closed loop is formed with PT research components, such as (i) data processing and fusion, (ii) analysis and modeling, in the middle. It is completed with offline and real-time functions connecting both supply and demand. Information on demand and supply dynamics enters PT research via AFC, APC, mobile crowdsensing data and AVL, GTFS data, respectively. The framework consists of five functionalities: service and operations planning, demand management, service control and management, (personalized) travel information and performance measurement. We briefly introduce these functionalities referring to the descriptions provided by Wilson et al. (2009).

- **Service and operation planning** consists of network and route design, frequency determination, and vehicle and crew scheduling. The offline analysis of PT data is significant for assessing the performance of routes and networks.

- **Demand management** relies on behavioral study of passengers and demand pattern analysis, which are both offline functions. The accumulation of such knowledge will help PT operators better design demand management strategies.

- **Service control and management** is inherently a real-time function which requires access to the current state of PT systems. It is an important function given that PT systems have the ability to communicate and process the data in real time to assist service controllers and managers. Critical information contains current location of vehicles, schedule deviation, and current onboard occupancy of each vehicles.

- **(Personalized) Travel information** can be either static, i.e. based on the scheduled or expected system performance, or dynamic, i.e. based on the actual state of the system. There can be pre-trip or en-route information to assist passengers'

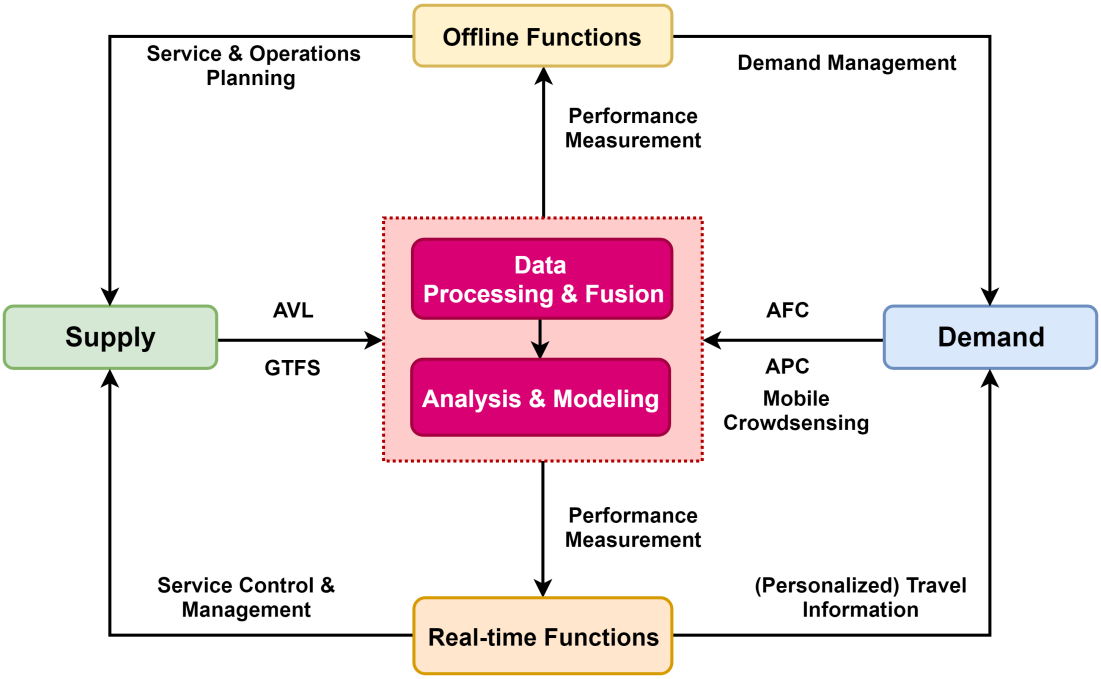


Figure 2.4: A framework for data-driven PT research, of which ultimate goal is to contribute to the improvement of PT planning and operations. This framework is a modified edition based on the ones presented by Koutsopoulos et al. (2017, 2019).

decision making. Currently, the trend is to provide personalized travel information based on the profile of individual travelers. Recommendation systems which take into account comfort (i.e., crowding level and number of transfers), travelers' preference on sustainability and multi-modality etc., are being researched and developed.

- **Performance measurement** as a real-time function is an important input into service control and management (e.g., in the form of dashboard capabilities), and as an offline function provides some of the data and information needed for service and operations planning. Critical information in performance monitoring and measurement includes on-time performance, passenger loads and service quality experienced by the customers.

The research that is going to be presented in the following chapters fits in this framework. We will cover both demand ans supply aspects, but will only be focused on offline functions.

Chapter 3

Onboard Occupancy Inference for Public Transport Vehicles

This chapter looks at how to obtain onboard occupancy of PT vehicles by integrating three different data sources, including AFC, AVL and GTFS. Although the problem looks minor, it has remained as a difficult task for PT operators due to technical and financial constraints. This chapter is therefore dedicated to addressing this problem. We first specifically identify the issues related to each and the combination of different data sources. Then based on this diagnosis, we propose a methodology for systematically addressing these issues, which results in desired profiles of PT vehicles with onboard occupancy and improved vehicle trajectories. We demonstrate the proposed methodology using the data from the PT system of The Hague, the Netherlands. The resulting profiles are visualized using space-time seat occupancy diagrams, which provides operators with a compact and powerful tool to intuitively examine the onboard crowding patterns over time and space. This visualization technique can help operators in timetable optimization, network and fleet scheduling, and sub-route service designing.

The chapter is organized as follows. Section 3.1 introduces the background and related literature. Section 3.2 describes the prevailing PT data sources, including AFC, AVL and GTFS data, along with related data issues. The methodology is described in section 3.3 with an overview and descriptions of all the steps as well as implementation. Section 3.4 presents the results, followed by the conclusions and discussion of future research directions in section 3.5.

This chapter is an edited version of the following article:

Luo, D., Bonnetain, L., Cats, O. & van Lint, H. (2018) Constructing spatiotemporal load profiles of transit vehicles with multiple data sources. *Transportation Research Record*, 2672(8), 175-186.

3.1 Introduction

Knowing the onboard load of vehicles is key to improve PT services from both planning and operational perspectives. However, obtaining such information for PT operators has remained a difficult task for a long time due to technical and financial constraints. Although manual surveys have often been utilized to estimate on-board passenger loads, such surveys are too costly to be conducted daily over all offered services, and are also subject to error and bias. Opportunities to change this situation, however, have emerged in the past years with the fast growing data richness in PT research and practice, including AVL (Moreira-Matias et al., 2015); AFC (Utsunomiya et al., 2006; Pelletier et al., 2011); and GTFS data (Wong, 2013). In many cities and regions around the world, PT demand and supply data has been continuously collected and managed with fine granularity, accuracy and spatiotemporal scale. Notwithstanding, it is not uncommon to see PT operators still struggle with the obtaining of some fundamentally important information, such as service utilization (i.e. passenger load). Data are often underutilized due to considerable deficiencies and shortcomings that can be frequently overlooked. To unlock their potential, it becomes necessary to develop sound techniques to overcome these issues and achieve valuable information, such as the PT vehicle load, by processing and integrating different data sources, including AFC, AVL and GTFS. The current study is therefore dedicated to this specific challenge.

To the best of our knowledge, few scientific studies and practical reports have attempted to address a similar problem. One of the main causes for this scarcity could be the rather limited access to multiple PT data sets from the same period by researchers. In many cases, only a single data source is available and the studies primarily developed methods to infer missing information. For example, Alfred Chu & Chapleau (2008) early on presented how spatiotemporal bus load profiles could be estimated based on AFC data only. In the absence of real bus trajectory information, they managed to estimate the spatiotemporal paths of vehicles by combining the first and last transaction times at each stop and corresponding timetable. Their work is one of the pioneering studies that revealed the power of AFC data on load profile construction. Sun et al. (2012) subsequently investigated a similar problem, however, in the context of a metro system. With only AFC data available (both tap-in and tap-off information recorded), they developed a methodology for estimating trains' trajectories and linked individual passenger rides to these trajectories, which results in a spatiotemporal density of metro vehicles. More recently, Moreira-Matias & Cats (2016) proposed a novel method for estimating on-board loads of buses using AVL data only. Passenger loads are built by applying machine learning algorithms to smoothen the load profile based on actual dwell time records. In addition, a web-based application to visualize bus load profiles, called BusViz, has also been developed based on the AFC data in Singapore (Anwar et al., 2016). Despite the progress on the visualization work, their approach to derive bus trajectories has several constrained assumptions. For instance, the arrival

time of a bus at a stop is equated to the earliest entry time of the first passenger who boards or alights at that stop, while the departure time is set equal to the greatest of the card entry times of passengers who board or alight at the stop. In order to improve this weakness, more advanced vehicle trajectory inference techniques based on AFC data only can be adopted and extended to address these limitations (e.g., Min et al., 2016; Zhou et al., 2017)

Although multiple PT data sets that are comparable among each other have become increasingly available to researchers, most research effort has focused on a selected number of fields, such as PT origin-destination estimation (Nassir et al., 2011; Gordon et al., 2013), travel time reliability analysis (Ma et al., 2015, 2017), and passenger assignment modeling in urban rail systems (Kusakabe et al., 2010; Zhao et al., 2017a; Hörcher et al., 2017; Zhu et al., 2017). Few existing studies have comprehensively examined how spatiotemporal load profiles of PT vehicles can be constructed using multiple data sources. This study is hence devoted to bridging this gap, which can benefit both researchers and practitioners. Our contribution is twofold, including specific identification of the issues pertaining to a single or a combination of data sets (AFC, AVL and GTFS), and the development of a methodology for addressing these issues and generating spatiotemporal load profiles of PT vehicles. The methodology consists of four steps through which raw data are processed and integrated to generate the passenger load profiles over space and time. These profiles allow service providers to analyze vehicle trajectories and demand patterns, and further investigate service utilization and the propagation of delays and crowding. The data collected from the urban PT network in The Hague, The Netherlands are utilized for demonstrating the methodology. A series of inference and matching steps are employed. This analysis results in profiles of vehicle trajectories and passenger loads which are further visualized through space-time occupancy graphs. Analogously to how space-time graphs of speed and flow enable traffic engineers to study spatiotemporal congestion patterns along routes in car traffic, these space-time occupancy graphs enable PT operators to study and inspect spatiotemporal on-board crowding patterns along PT service lines.

3.2 Identifying data issues

Several issues pertaining to a single or a combination of data sets can be identified and need to be resolved for the current application, i.e., constructing the spatiotemporal load profiles of PT vehicles. These issues are summarized and presented in Figure 3.1 in relation to their sources. Basically, issues specific to each individual data set are illustrated in the respective oval, whereas issues that arise when two or more data sources are combined are positioned at their intersections. The following issues have been identified:

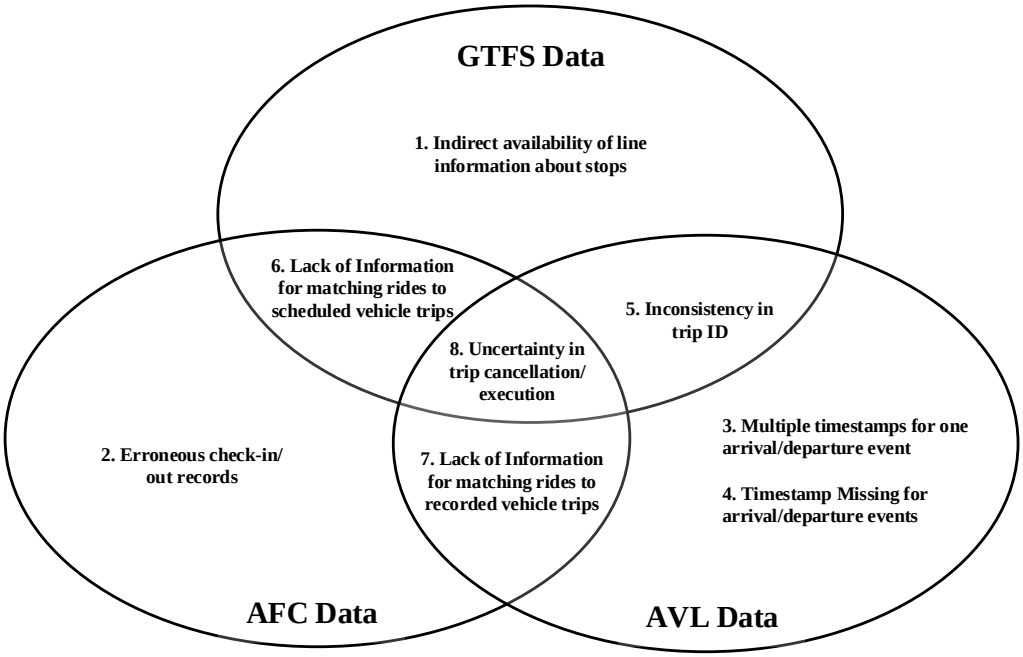


Figure 3.1: Identification of the issues pertaining to a single or a combination of data sets for constructing the spatiotemporal load profiles of PT vehicles. Issues specific to each individual data set are illustrated in respective ovals, whereas issues that arise when two or more data sources are combined are positioned at their intersections.

1) **Indirect availability of line information about stops**
   Information in GTFS data is stored in a trip-based manner, meaning that what can be directly obtained from it are only individual vehicle trips that consist of all the stops per trip. Complete stop sets of service lines, which are crucial for subsequent analytics, are not directly available from the GTFS data.

2) **Erroneous AFC check-in/out records**
   The arrival time at any given stop cannot be earlier than the departure time from the last stop.

3) **Multiple timestamps for one arrival/departure event**
   Multiple timestamps can be occasionally found for one arrival or departure event in the raw AVL data set. It is unclear why this happens, but it jeopardizes the global consistency of vehicle trajectories.

4) **Timestamps missing for one arrival/departure event**
   Arrival and/or departure timestamps at a stop can be missing. The size of missing timestamps ranges from one event (arrival/departure) at a stop to an entire trip. Note that issue #3 and this one can happen to the same stop (e.g., two arrival events, missing departure event), which makes the issue even worse.

5) **Inconsistency in trip ID**
   The vehicle trip ID indices in AVL and GTFS data do not always match. This

inconsistency causes problems in matching trip and trajectory when combining GTFS (scheduled trajectories) and AVL (recorded trajectories).

6) **Lack of Information for matching rides to scheduled vehicle trips**
Since the AFC data set used in this study does not contain vehicle trip ID, it is impossible to directly match individual rides to scheduled vehicle trips that are extracted from the GTFS data.

7) **Lack of Information for matching rides to recorded vehicle trips**
The same issue as #6 holds for this situation too. There is no direct way to match rides to the recorded vehicle trips from AVL data.

8) **Uncertainty in trip cancellation/execution**
The GTFS data contain all scheduled trips of a day. However, this does not provide conclusive evidence that all these trips are indeed executed. In many cases, the number of trips found in the AVL data set is smaller than the scheduled number of trips. It is uncertain whether this is a result of trip cancellation or AVL system malfunction without any additional information. AFC data may be used here to settle the discrepancy.

These issues are illustrated for a given day and line in Figure 3.2, which visualizes the recorded trajectories from AVL data; scheduled trajectories from GTFS data; and check-in/out records. 3.2a first displays all the recorded trajectories from the AVL data. There are many gaps in this plot, which indicates that there is either a missing timestamp or multiple timestamps for the arrival or departure event at that stop. Figure 3.2b adds the layer of all scheduled trajectories (blue lines) underneath the recorded ones (red lines). It can be observed that overall vehicle trips adhere to the timetable very well. Next, check-out (blue circle points) activities are added in Figure 3.2c. An important finding from this plot is that when there is a trajectory gap, check-in/out activities also do not exist, or are very sparse, which implies that in the case where the arrival timestamp is missing but departure has at least one timestamp, the vehicle probably drives through the stop without serving passengers. Figure 3.2d displays a zoom-in plot to allow for a more detailed inspection. Check-in activities are clustered close to the vehicle arrival time, unlike check-out activities, because it is customary for passengers to check out in the segment directly upstream of the alighting stop.

3.3 Methodology

In this section, a methodology for constructing the spatiotemporal loads of PT vehicles based on aforementioned data sources is described. An overview is first provided, followed by subsections dedicated to each step. The final subsection describes how this was implemented.

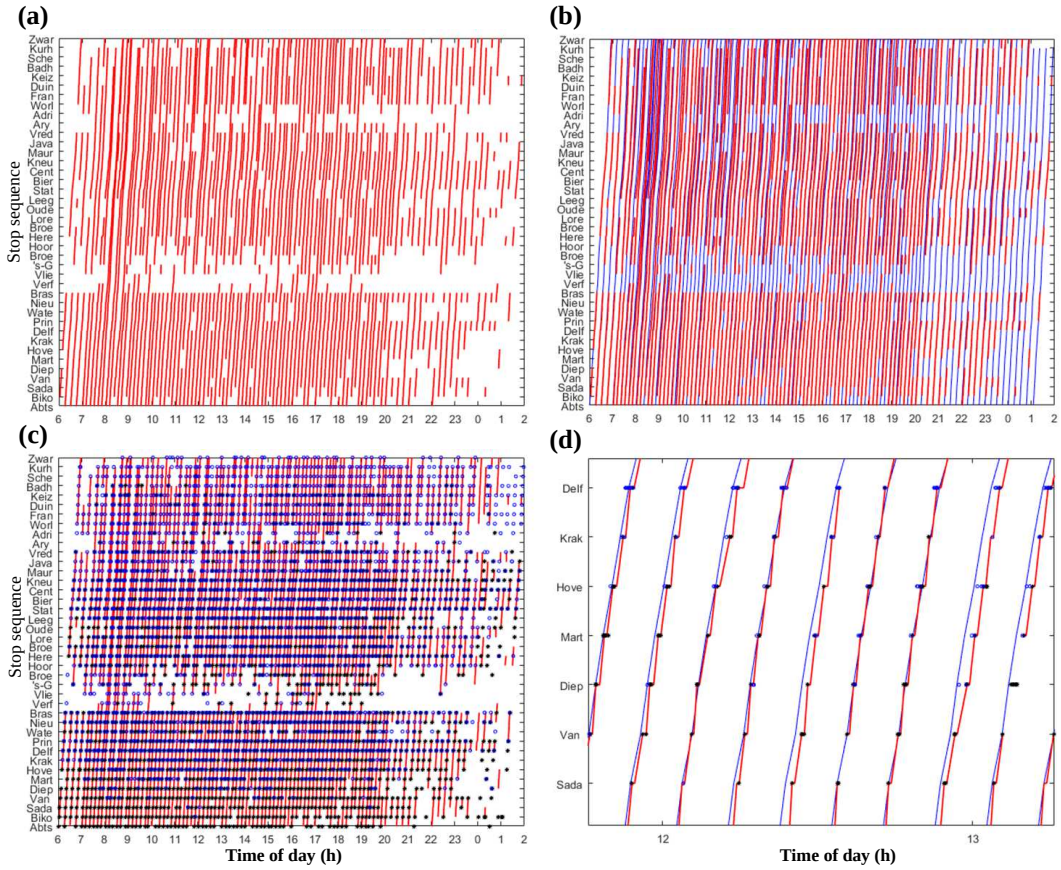


Figure 3.2: Visualization on how different data sources characterize a full-day service. The example pertains to the operations of tram line 1 (Delft Tanthof to Scheveningen Noorderstrand) on March 5, 2015. (a) Recorded trajectories (red lines) obtained from the AVL data set; (b) Recorded trajectories (red lines) on top of all the scheduled trajectories obtained from the GTFS data set; (c) Recorded trajectories (red lines) on top of all the check-in (black star points) and check-out (blue circle points) activities; (d) Zooming-in for a selected hour (12-13) of the data presented in (c).

3.3.1 Overview

The methodology consists of four steps as shown in Figure 3.3. Raw data from three independent sources is used throughout the four steps. All three individual data sets are initially stored in separate databases. In step 1, the raw data is first obtained from the databases with all the information restructured at a daily level. By doing so, the subsequent workflow is made more computationally efficient. Issues #1 and #2 are addressed at the first step, resulting in data files respectively containing passenger rides; recorded vehicle trajectories, and scheduled vehicle trajectories. The integration of recorded and scheduled vehicle trajectories is then performed at the second step. Issue #5 is solved resulting in data files that contain both scheduled and recorded vehicle trajectories. In step 3, passenger rides are matched with vehicle trips and trajectories (solving issues #6 and #7). In the last step, all scheduled trips are first labeled either

"canceled" or "executed" based on the validation results. The data files from step 3 are used to perform this validation task, addressing issue #8. Finally, vehicle trajectories of validated trips are corrected by fusing multiple types of information. Consequently, all trajectories have complete trajectory and load information and are globally consistent. Issues #3 and #4 are hence solved in this final step. This sequential method decomposes the process into small sub-tasks with each step solving one or several of the identified issues.

3.3.2 Step 1: pre-processing data

All three types of data are preprocessed in the first step at a daily level. For AFC data, single transaction records are first linked to generate individual passenger rides with both check-in and check-out information (stops and times). Erroneous rides with unrealistic travel time and origin or alighting stops are identified and removed in this process. In this case, the travel time threshold was set to 90 min, which exceeds the maximal travel time between any pair of stops in the case study network. In addition, the indices of stops were also transformed to be consistent with those of the AVL and GTFS data.

Every single scheduled trajectory - characterized by the arrival and departure time at every stop along a line - is extracted from the GTFS data. This process is not straightforward due to the fact that the GTFS standard does not contain direct information about regular stop sequences of individual PT routes. Designed originally for the purpose of route planning, the GTFS data makes it quite handy to obtain arrival and departure times at each stop of individual trips by storing information based on trips. A trip is recorded in the trip.txt file and further detailed in the stop_times.txt file with its sequence of stops. All trips, including sub-lines and partial trips with some stops skipped, are thus easily stored in the data. However, this becomes an obstacle when we want to obtain the most regular and fullest stop sequence of PT routes. To overcome this problem, a brute-force approach is adopted. All trips of a PT route from a normal working day are scanned in order to acquire the complete set and right sequence of stops on this line and this direction. This is a straightforward yet effective solution to this problem. AVL data, however, does not need to be much processed since the information is already organized based on stop sequences of trips. If there is a missing record, a "Not A Number" label is added.

3.3.3 Step 2: matching trips in GTFS and AVL

This step is dedicated to matching all the trips recorded in the AVL data set to all the scheduled trips contained in the GTFS data. Ideally, the two data sources should share the same trip ID indexing scheme so that the matching is very straightforward.

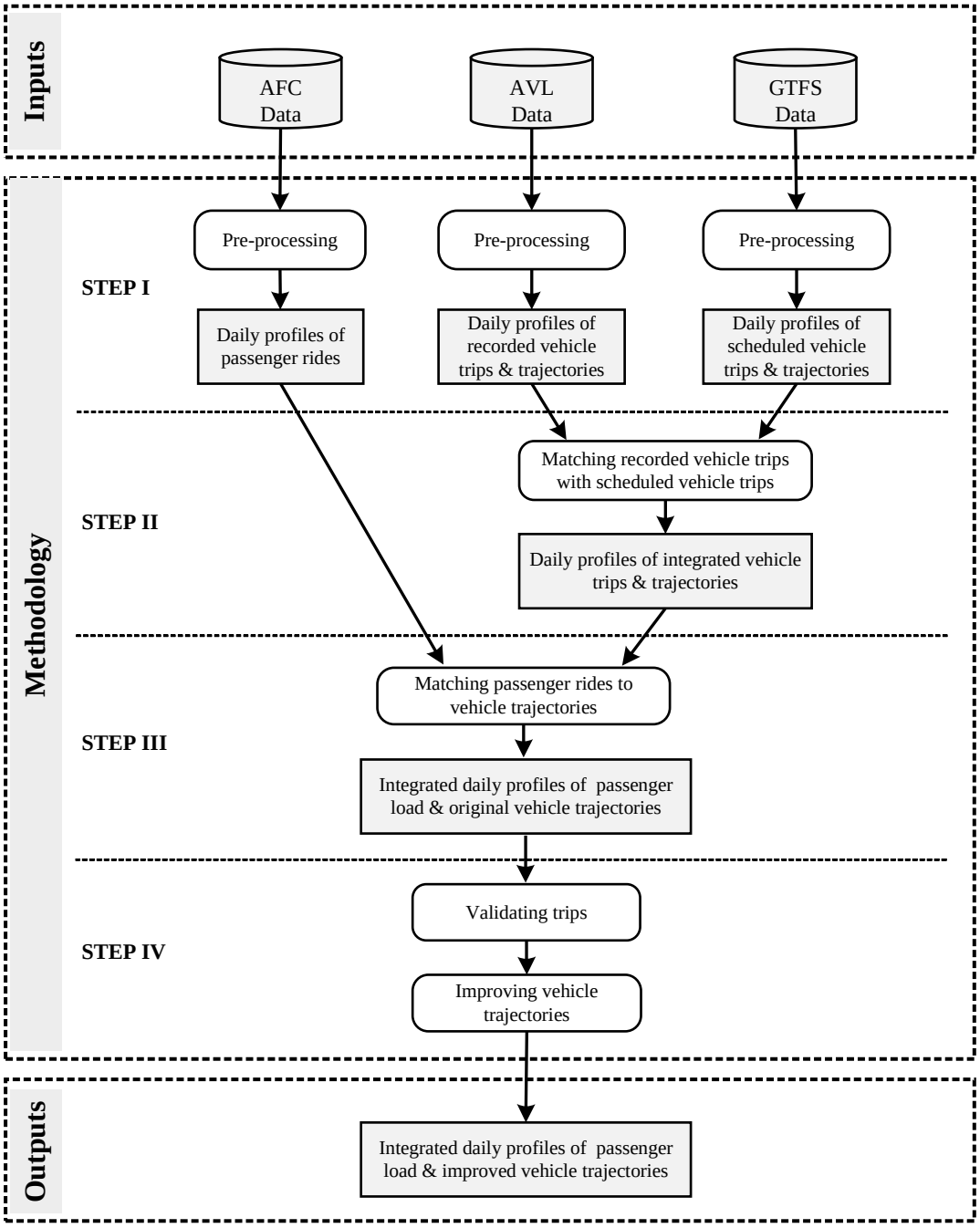


Figure 3.3: Overview of the four-step methodology. Inputs are raw information from individual data sets, and the final outputs are integrated profiles containing vehicle trajectories with passenger loads.

However, inconsistencies do exist as illustrated in the previous section. To address this issue, the recorded arrival time as well as the delay at stops from the AVL data is used to compute the probable scheduled arrival times as follows:

$$\tilde{\pi}^{s}_{n,k} = \pi^{r}_{n,k} - d^{r}_{n,k} \tag{3.1}$$

where $\tilde{\pi}^s_{n,k}$, and $\pi^r_{n,k}$, respectively, denote the probable scheduled and recorded arrival times of trip n at stop k. $d^r_{n,k}$ denotes the recorded delay of trip n at stop k from the AVL data. The scheduled trip that has the closest arrival time at a stop to this "estimated" scheduled arrival time $\pi^r_{n,k}$ is then found, and its trip ID from the GTFS data is temporarily labeled to this stop visit. After applying this process to all the stops of this trip, the GTFS trip ID that has been most frequently labeled is adopted and assigned to the entire trip. The matching of the recorded trips to scheduled trips is performed so that those trips that cannot be found in the AVL data will be later checked to assess whether they were really executed by taking the AFC data into consideration. In addition, headways based on AVL and GTFS are also computed and added to the trajectory profiles at the end of this step.

3.3.4 Step 3: matching passenger rides to vehicle trajectories

The objective of this step is to match all individual passenger rides to the vehicle trips that these passengers traveled with. However, since the trip ID information is missing in the current AFC data set, a trip ID inference algorithm for all the rides is first developed as shown in Figure 3.4.

Let $t^{in}_{i,k}$ denote the check-in time of passenger i at stop k. Let $\pi^r_{n,k}$ and $\pi^s_{n,k}$, respectively, denote the recorded and scheduled arrival times of trip n at stop k. Essentially, the algorithm attempts to find the trip ID for a single ride by the i-th passenger so that his or her check-in time at the stop k, $t^{in}_{i,k}$, is closest to the vehicle arrival time $\pi^r_{n,k}$ and $\pi^s_{n,k}$ at the very same stop k along the trip n. Recorded arrival times π^r are used as the major benchmark because delays can introduce a significant bias when performing such inference. If the condition shown below is satisfied, then this ride was labeled with the trip ID of trip n.

$$\pi^r_{n,k} - \varepsilon^- < t^{in}_{i,k} \leqslant \pi^r_{n,k} + \varepsilon^+ \tag{3.2}$$

where ε^- and ε^+, respectively, represent the lower and upper bounds of the searching time window. In this case study, ε^- and ε^+ were empirically set to be 20 and 50 seconds, respectively after scrutinizing the data. When the recorded arrival time of a trip at this stop π^r, nevertheless, is missing or has multiple values, the scheduled arrival time π^s is then employed with a larger time window shown as below:

$$\pi^s_{n,k} - \frac{h^s_{n,k}}{2} < t^{in}_{i,k} \leqslant \pi^s_{n,k} + \frac{h^s_{n+1,k}}{2} \tag{3.3}$$

where $h^s_{n,k}$ denotes the scheduled headway between the current trip n and previous trip $n-1$ at stop k. Finally, if the inference based on the recorded arrival time do not yield a result, the result based on the scheduled arrival time will be adopted.

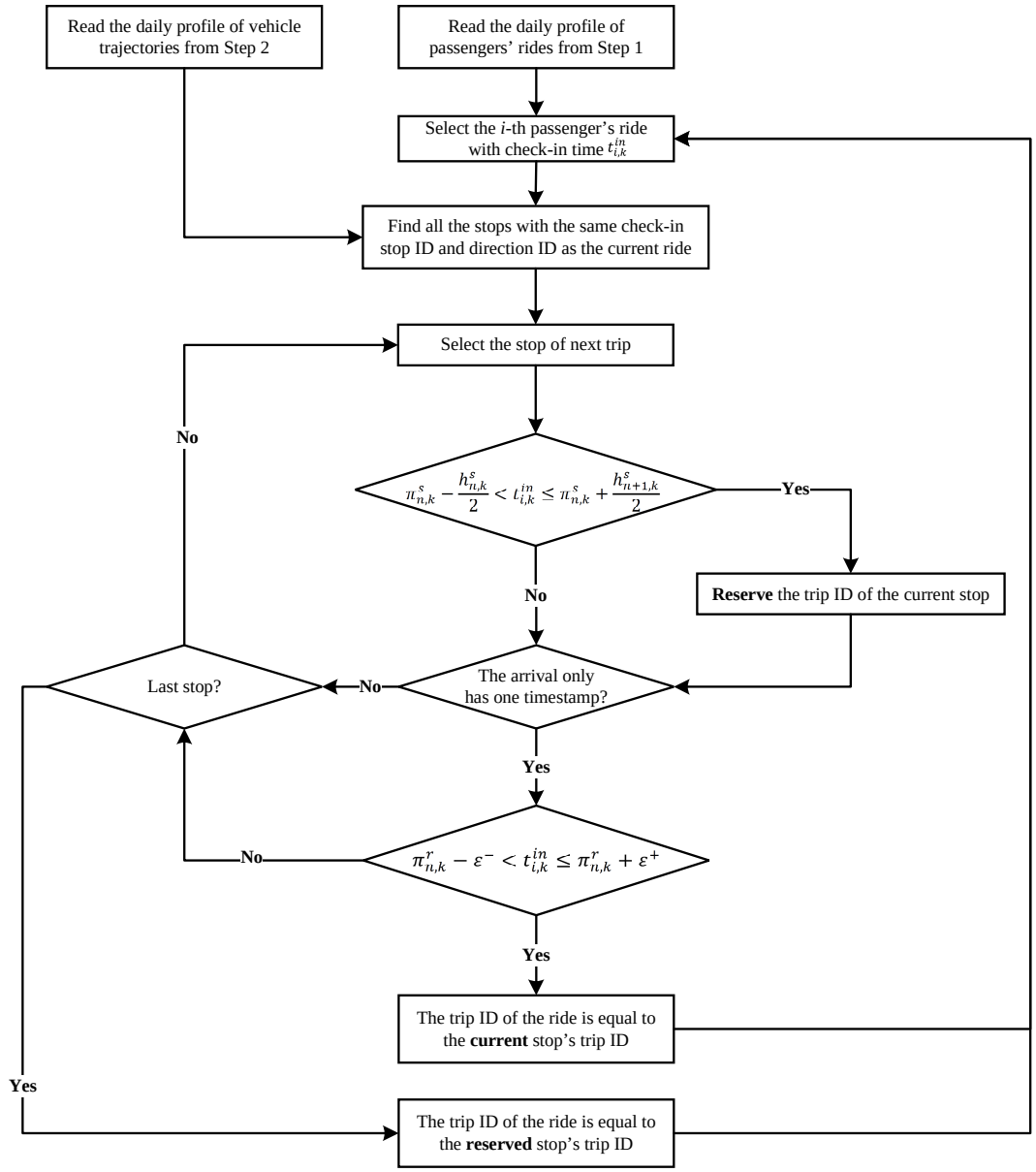


Figure 3.4: Algorithm for inferring the trip ID of individual passenger rides.

3.3.5 Step 4: improving vehicle trajectories

The last step is dedicated to the correction of vehicle trajectory profiles based on the results from step 3. It attempts to resolve the issues identified in the AVL data (#3, #4 and #8). The central idea is that the integrated information from three different data sources at the same detail level, including the recorded (AVL) and scheduled (GTFS) arrival/departure times, as well as check-in/out times at stops(AFC), will allow us to (1) infer whether a vehicle trip was indeed executed, and (2) restore actual vehicle trajectories to the maximal extent. In the current study, practical solutions consisting of a number of rules were developed and applied.

To infer whether a trip is executed or canceled in reality, a series of rules are proposed

based on the practical investigation. The inference takes into account (1) whether there is sufficient information about recorded timestamps for a trip; and (2) whether there are enough check-in activities that can be reasonably associated with this trip. Eventually, all scheduled trips are labeled as either "executed" or "canceled".

- **Rule 1:** The arrival time cannot be later than the departure time at a given stop.
- **Rule 2:** The arrival time at any given stop cannot be earlier than the departure time from the last stop.
- **Assumption 1:** When only a timestamp for the departure is available, the arrival time should be equal to the departure time because the vehicle presumably did not have to serve this stop.

By applying these rules and this assumption, it is ensured that the vehicle trajectory of a trip is globally consistent in a sense that vehicles can never move backwards. The assumption is made based on the practical investigation into the data. As Figure 3.2c illustrates, when there is a gap in the recorded trajectory, very few check-in/out activities can be spotted. Therefore, it can be safely assumed that vehicles skipped the stops when only departure times are logged in the database.

3.3.6 Implementation

All three raw data sets are stored in a PostgreSQL 9.3 database. A series of indices on date, line ID, stop ID, etc. were created to improve the SQL query performance. All of the abovementioned steps were coded in MATLAB.

3.4 Results

For simplicity and tractability, only the results of tram line 1 for the entire month of March 2015 are described in this section. Line 1 connects Delft, a mid-size old university city to The Hague, the main city in its urban agglomeration, serving 41 stops per direction, including three major train stations. The service is frequent with 208 trips on a normal weekday and up to 8 trams per hour in the peak on each direction. More than 670,000 passenger rides are recorded for line 1 in the AFC database over the case study period. After applying the pre-processing to these rides, around 0.03% (around 200) rides were discarded due to the data issues described above.

Figure 3.5a shows the result of vehicle trip ID inference for passenger rides. For most of the days, over 90% of the rides' trip ID can be successfully inferred based on the recorded trajectories owning to the relatively high quality of AVL data on these days. Two particular days, i.e., the 6th and 18th of March, are noticeable due to their low

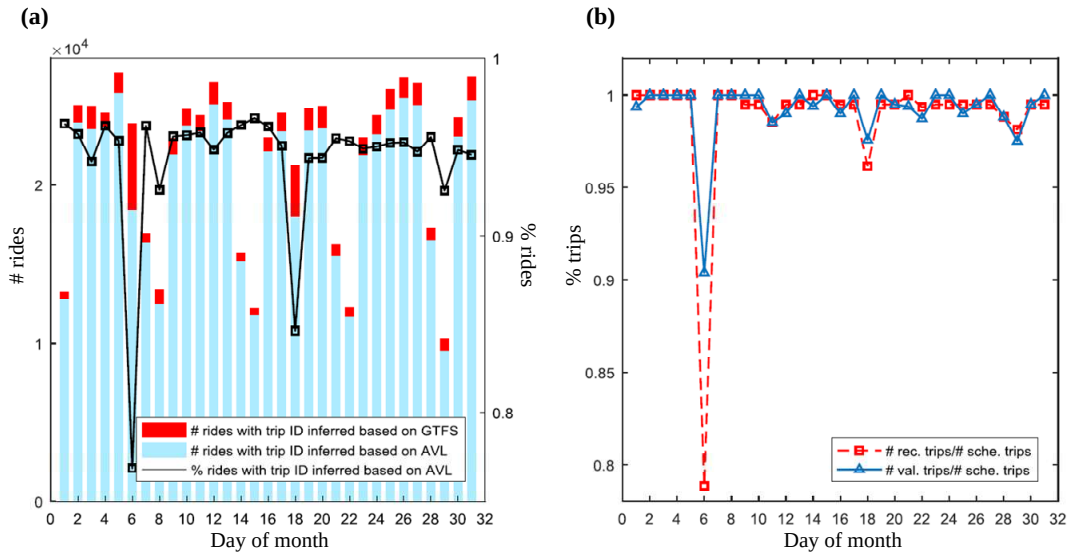


Figure 3.5: Results for trip ID inference of rides and trip validation. (a) Illustration of trip ID inference for rides that are based on recorded trajectories (AVL) and scheduled trajectories (GTFS), respectively. The line shows the percentage of rides of which trip IDs are inferred based on the recorded trajectories (AVL); (b) Comparison among the numbers of scheduled trips, recorded trips and validated trips.

AVL-based inference percentage. This is arguably caused due to the significant loss of AVL data on those two days, which implies that many inferences rely instead on just the scheduled trajectories.

The percentages of recorded trips in the AVL data and the executed trips that are eventually validated compared to the total number of scheduled trips are calculated and displayed in Figure 3.5b. Both directions are considered at the same time. It can be seen that on many days, not all scheduled trips were executed. A significant plunge appears on the 6th of March, where fewer than 80% of scheduled trips are recorded in the AVL data, whereas over 90% trips could be validated when also using information from check-in/out activities on that day. In addition, the 18th and 29th also yield results that are clearly worse than the average. As a result of inconsistent and incomplete trajectories from the AVL data, trip ID inference of rides on these days become more unreliable, resulting in a stronger reliance on GTFS-based inference as shown in Figure 3.5a for March 18th and 29th.

The final output profiles are visualized by plotting so-called space-time vehicle seat occupancy graphs (Figure 3.6). These seat occupancy graphs relate the vehicle occupancy to the seating capacity. In the color schemes in Figure 3.6, 100% occupancy means that the number of passengers on board is equal to the number of seats (76 for all the vehicles running on line 1). Significant crowdedness is thus easily identified when the seat occupancy is higher than 100% (warmer color). The upper bound is set to 200%, corresponding to the maximal vehicle capacity (around 150 people). We want to emphasize that this visualization technique has great potential for decision support

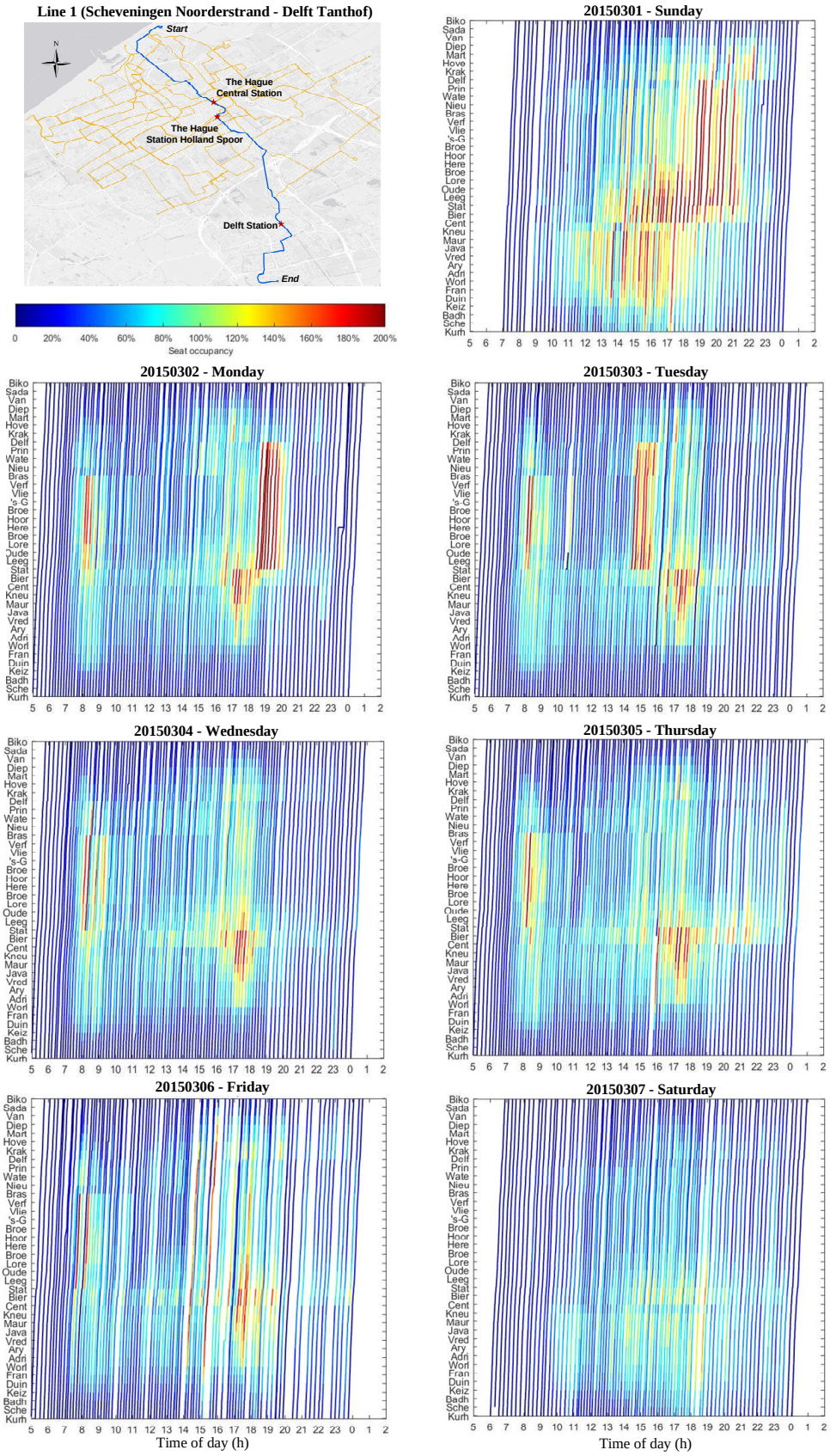


Figure 3.6: Illustrations of spatiotemporal seat occupancy of line 1 from Scheveningen Noorderstrand to Delft Tanthof over the first week of March 2015.

in PT planning and operations, ranging from timetable optimization; network and fleet scheduling; designing sub-services running over partial lines, to name just a few. The graphs provide a single, information-rich and intuitive global view of service quality.

Although detailed analyses are beyond the scope of this exploratory study, example visualizations presented in Figure 3.6 (line 1 from Scheveningen Noorderstrand to Delft Tanthof over the first week of March 2015. March 1st is a Sunday) allow for identification of crowdedness over space and time. Overall, similar crowding patterns can be recognized on weekdays with clearly visible morning and afternoon peak-hour flows, while there are more variations during weekends. For instance, severe on-board crowding always occurs from Station Hollands Spoor to Badhuiskade in the morning, and from Vredespaleis to Station Hollands Spoor in the afternoon during this week.

3.5 Conclusion

Obtaining on-board load profiles of PT vehicles has remained a difficult task for operators in the past years due to technical and financial constraints. In this chapter, a new methodology for constructing such profiles with multiple PT data sources is presented, including AFC, AVL, and GTFS. Difficulties of utilizing these data are discussed with the issues arising from a single or a combination of data sets specifically identified. The methodology consists of four steps through which raw information from individual data sources is processed and corrected. The output profiles can convey integrated information regarding both vehicle trajectories and passenger demand on a large spatiotemporal scale. The methodology is demonstrated with the data collected from the urban PT system in The Hague, The Netherlands. A key output is so-called space-time seat occupancy graphs, which provides operators with a compact and powerful reference to intuitively examine the on-board crowding patterns over time and space, thus helping to improve their services, such as timetable optimization; network and fleet scheduling; designing sub-services running over partial lines, etc.

The contribution of this chapter is twofold. Firstly, we endeavor to integrate different PT data sources for obtaining state estimations for passenger loads. In this process, the issues related to each and the combination of different data sets, namely AFC, AVL and GTFS, are specifically identified. Although based on data available in the Dutch context, most of their properties are universal, and our way of presenting all the issues can be beneficial for researchers and practitioners with different data formats but similar difficulties. Secondly, a methodology that solves these issues in a sequential manner is described and yields service profiles containing both vehicle trajectories and passenger loads. The complexity of approaches and algorithms in each step can vary depending on the availability of information.

Future research can be performed in the following directions. First, the inference techniques for matching passenger rides with vehicle trajectories and correcting vehicles' trajectories can be improved. This could be realized by replacing the current simple

rule-based algorithms with more advanced probabilistic or machine learning models. Inspirations on how arrival times can be derived from AFC data records can be obtained from the study by Zhou et al. (2017). Second, network-level dynamics and regularities of passengers, vehicles, and even their interactions can be thoroughly investigated provided with the resulting vehicle profiles. Such studies on day-to-day regularity of delay and crowding patterns at a network level can be beneficial for the planning of PT frequency and timetable.

Chapter 4

Principal Component Analysis of Passenger Flows

This chapter deals with the issue of high-dimensionality in large-scale passenger flows. This issue makes it challenging to analyze and model passenger flow dynamics in a large spatiotemporal scale. We address this challenge by applying principal component analysis (PCA), a popular dimensionality reduction technique. We first show how the matrix of multivariate time series of passenger flows can be constructed, and then specify how such high-dimensional flow matrix can be transformed to a lower dimensional space using PCA. To demonstrate the methodology, we use the AFC data from the metro system of Shenzhen, China in the case study. The results show that a great amount of variance contained in the original data can be retained with much fewer principal components (8 for 90% and 29 for 95%). Our further analysis on the temporal stability of the flow structure in this lower dimension shows that the principal components obtained from historical data are capable of approximating the profile of future data when there are no non-recurrent or special events. This chapter contributes to understanding large-scale spatiotemporal PT flow patterns.

The chapter is structured as follows. Section 4.1 provides the research background and motivation. Section 4.2 explains the methodology about how PCA can be applied to multivariate flows. Then the details of the case study are described in section 4.3, followed by the presentation of results and discussion in section 4.4. Section 4.5 presents the conclusion with suggestions for future research directions.

This chapter is an edited version of the following article:

Luo, D., Cats, O. & van Lint, H. (2017) Analysis of network-wide transit passenger flows based on principal component analysis. In *proceedings of the 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 744-749.

4.1 Introduction

PT systems have been rapidly developed in many places as a solution to mobility and environmental problems, especially in metropolitan areas with dense population and limited road resources. These PT systems, mostly comprised of lines of buses, trams and metros, are being used by a great number of travelers every day for all kinds of activities. As many PT systems are still expanding and attracting more travelers, it becomes imperative for PT researchers, managers and operators to gain more knowledge on such complex systems, which can largely benefit the development of advanced PT management tools. For example, one of the significant tasks is to understand the dynamics of passenger flows in PT systems in order to facilitate advanced PT fleet and demand management. This particular task, which was until recently hindered by limited amount of flow observations, has now been enabled thanks to the vast amount of smart card data collected by AFC systems Pelletier et al. (2011). Since individual travelers' trips are passively recorded while they travel, the information contained in such data is very complete and characterized by fine granularity. It hence provides both practitioners and researchers with a precious chance to investigate PT mobility and flow dynamics in depth.

Numerous studies which leverage smart cart data to unravel the spatial-temporal patterns of PT trips, urban mobility and travelers' behaviors have been published, such as Sun et al. (2013). Besides, researchers have also attempted to measure the variability of mobility patterns (Zhong et al., 2015) and identify urban activity centers or clusters (Cats et al., 2015) based on passenger flows obtained from smart card data. These existing studies succeeded in strengthening our understanding of urban mobility and PT systems, but a limited number of them were found to shed light upon passenger flow dynamics from a multivariate perspective, which means to deal with the high dimensionality of such flow data. Due to the existence of the so-called "curse of dimensionality" (Verleysen & François, 2011), the development of a series of important PT applications, such as network-wide flow modeling and prediction, might be hindered without sufficient insight into these multivariate passenger flows. It becomes substantially difficult to find effective and intuitive solutions in a high-dimensional space while dealing with complex systems like a PT network which consists of multiple lines and hundreds of stations. As argued by Verleysen & François (2011), this particular difficulty results from the conjunction of two effects. Firstly, some geometrical properties of high-dimensional spaces are counter-intuitive and different from what can be observed in 2- or 3-dimensional spaces. Secondly, data analysis tools are usually designed in low-dimensional spaces with intuition.

Given the research gap and difficulty described above, this study is aimed at performing a multivariate analysis of PT passenger flows based on a well-known dimensionality reduction technique, PCA. We detail how a one-month smart cart data set from the Shenzhen metro system is transformed to multivariate time series of flows and how PCA is performed on such time series. The results of PCA, including the low dimen-

sionality of flows, features of principal components (PCs), approximation of original flows, and temporal stability of flow structure, are explicitly presented and analyzed, providing an insight into the underlying structure of flow dynamics within a complex PT network. Overall, this study contributes to the development of multivariate analysis on PT passenger flows, and shows the potential of incorporating PCA into promising applications, such as anomaly detection and short-term forecasting.

4.2 Methodology

PCA was initially proposed to describe the variation of a set of uncorrelated variables in a multivariate data set (Pearson, 1901; Hotelling, 1933). So far it has been extensively used for various tasks, such as dimensionality reduction, factor analysis, feature extraction, and lossy data compression. In the field of traffic and transportation, for example, PCA was integrated into dynamic OD estimation and prediction in order to overcome the computational problem caused by high-dimensional OD matrix data (Djukic et al., 2012). It was also leveraged to analyze travelers' longitudinal behavior by extracting the so-called eigen-sequences (Goulet Langlois et al., 2016), to name a few.

The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. PCA achieves this target by projecting the observations onto a new set of axes which are called the PCs. Each PC has the property that it points in the direction of maximum variance remaining in the data, given the variance already accounted for in the preceding components. As such, the first PC captures the total energy of the original data to the maximal degree possible on a single axis. The following PCs then capture the maximum residual energy among the remaining orthogonal directions. In this sense, the PCs are ordered by the amount of energy in the data they capture.

It has been shown that PCA can be used as an effective tool to analyze whole-network traffic flows which are essentially high-dimensional multivariate time series (Lakhina et al., 2004). By performing PCA on the flow data, a smaller number of dimensions can be found and leveraged to well approximate original high-dimensional data. Let $\mathbf{X}$ denote a matrix of multivariate flow time series as equation (4.1) shows. Each column i of $\mathbf{X}$ denotes a single flow variable, while each row j represents an observation of all flow variables at time j. This yields a $t \times p$ matrix $\mathbf{X}$, where t represents the total number of time instances and p represents the total number of flow variables.

$$\mathbf{X} = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_p(1) \\ x_1(2) & x_2(2) & \dots & x_p(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t) & x_2(t) & \dots & x_p(t) \end{bmatrix} \tag{4.1}$$

As shown in equation (4.2), obtaining all the PCs of $\mathbf{X}$ is actually equivalent to calculating the eigenvectors of $\mathbf{X^T X}$ which is a measure of the covariance between flows.

$$\mathbf{X^T X v_i} = \lambda_i \mathbf{v_i} \tag{4.2}$$

where λ_i is the eigenvalue corresponding to eigenvector $\mathbf{v_i}$ $(p \times 1)$ and the number of eigenvalues/eigenvectors is equal to the number of variables p. In fact, the eigenvalue λ_i indicates how much variance of the original data is explained by the dimension i specified by eigenvector $\mathbf{v_i}$.

$$\mathrm{Var}(\mathbf{v_i}^\mathbf{T} \mathbf{X}) = \lambda_i \tag{4.3}$$

Arranging all the eigenvalues in a descending order ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$), the first PC is thus the eigenvector which corresponds to the largest eigenvalue since it accounts for the greatest variance in the entire data.

By mapping the original data onto the derived principal component space, it can be seen that the contribution of dimension i (the i-th PC) as a function of time is given by $\mathbf{X v_i}$. Normalizing this vector to unit length with λ_i as shown in equation (4.4), we obtain a $t \times 1$ vector $\mathbf{u_i}$ which contains the information of temporal variation along the i-th PC. As a matter of fact, the vector $\mathbf{u_i}$ captures the temporal variation common to all flows along this dimension (PC). The set of vectors $[\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_p}]$, which are perpendicular, can thus be referred to as the **eigen-flows** of $\mathbf{X}$.

$$\mathbf{u_i} = \frac{\mathbf{X v_i}}{\sqrt{\lambda_i}} \tag{4.4}$$

Let $\mathbf{V}$ denote a $p \times p$ matrix consisting of all the PCs $[\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_p}]$ which are arranged in order. The first column $\mathbf{v_1}$ refers to the first PC, and so on. Let $\mathbf{U}$ denote a $t \times p$ matrix of which column i is $\mathbf{u_i}$. Consequently, each individual flow $\mathbf{X_i}$ can be written as:

$$\frac{\mathbf{X_i}}{\sqrt{\lambda_i}} = \mathbf{U}(\mathbf{V^T})_\mathbf{i} \tag{4.5}$$

where $\mathbf{X_i}$ is the time series of i-th flow and $(\mathbf{V^T})_\mathbf{i}$ is the i-th row of $\mathbf{V}$. This equation indicates that each flow $\mathbf{X_i}$ is essentially a linear combination of the eigen-flows with weights specified by $(\mathbf{V^T})_\mathbf{i}$.

By selecting the first r $(r \leq p)$ eigenvectors with largest eigenvalues, the information contained in original data $\mathbf{X}$ can then be effectively transformed onto a r-dimensional subspace of $\mathbb{R}^p$. It is shown in equation (4.6) how the approximation can be done.

$$\mathbf{X}' \approx \sum_{i=1}^{r} \sqrt{\lambda_i} \mathbf{u_i} \mathbf{v_i^T} \tag{4.6}$$

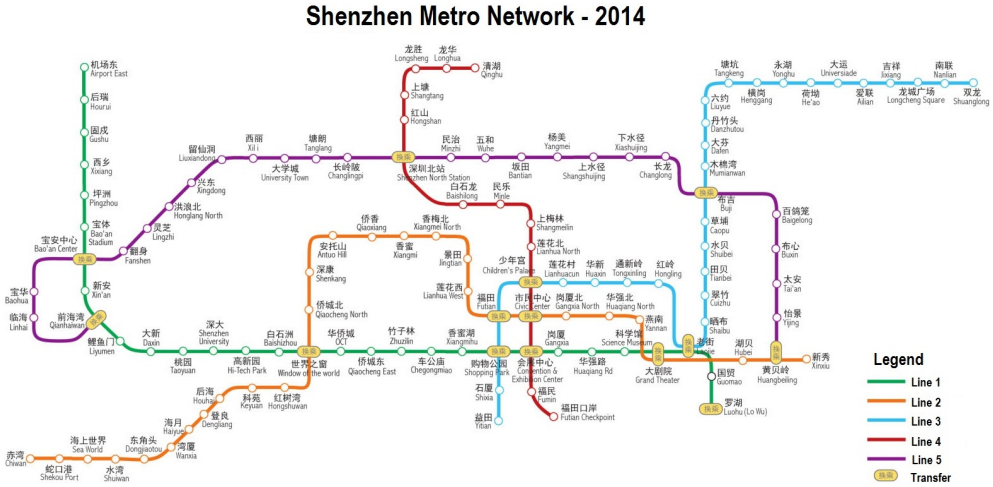


Figure 4.1: An illustration of Shenzhen metro network (2014).

4.3 Case study on the Shenzhen metro system

The 2014 Shenzhen metro system was employed for the case study. Shenzhen is one of the largest cities in China with a metropolitan area population of over 18 million. By the end of 2014, there were in total five lines in operation with 118 stations. An illustration of the network is shown in Fig. 4.1, where five lines are represented with different colors, and all transfer stations are highlighted with yellow marks. As a major transportation service in Shenzhen, the metro system accounts for approximately one third of the total public transport passenger traffic, which results in complex passenger flow dynamics over time. It is therefore a significant task to understand these flow dynamics for achieving better system operations and management.

AFC system was employed by the Shenzhen metro system and passengers could not travel without using a smart card. Moreover, tapping is required for both entry and exit activities because the fare was collected using a distance-based scheme. As a result, complete travel information of individuals except for transfer activities were recorded in the database. A typical record includes the time-invariant anonymous card ID, metro station ID, transaction timestamp, and transaction type (21 for tap-in and 22 for tap-out). The data set used for this study contains 139,646,884 records, covering the whole period of December of 2014. The period includes 23 normal weekdays and four weekends.

4.3.1 Constructing entry and exit flow profiles

We analyzed entry and exit flows at the same time. Raw data were transformed from a per-user basis to a per-station basis with time discretization. Time series of both entry and exit flows for each station with a time interval of 5 minutes were constructed. In line with the operational time of the metro system, the time horizon considered in this study was from 6 AM to 11 PM each day (17 hours). The total number of

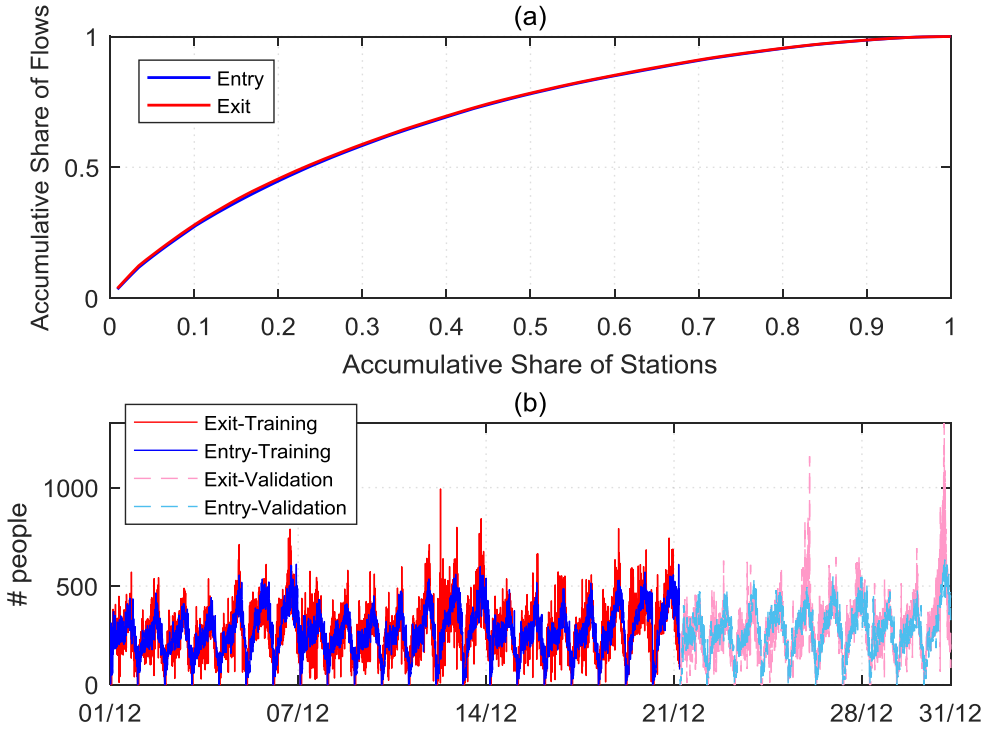


Figure 4.2: Illustrations of flow profiles. (a) Cumulative distribution function plots of entry and exit flows; (b) A typical example of entry and exit flow time series of Shenzhen metro station (Luohu station).

measurements for each flow variable over the entire period is then 6324 ($= 12 \times 17 \times 31$). Cumulative distribution function (CDF) plots of these flow profiles are shown in Fig. 4.2a. It can be understood from the diagrams that, in this case, flows are accounted for by a relatively large percentage of stations rather than only a few. This feature will be reflected in the following analysis.

4.3.2 Preparing training and validation sets

In order to examine whether the structure of entry and exit flows is temporally stable, the entire data set was divided into training part and validation part, with the former containing the data from the first three weeks (2014-12-01 to 2014-12-21) and the latter from the rest of the days (2014-12-22 to 2014-12-31). Note that Christmas day and New Year Eve are included in the validation set. The main motivation is to know whether the decomposition of entry and exit flows into eigen-flows, as determined by the set of PCs, is useful for analyzing data that are not part of the input to the PCA procedure. This is crucial for applications like forecasting. By using the training data alone, we obtained a 4284×236 flow matrix following equation (4.1). With p equal to 236, the first half of the columns were filled with entry flows while the second half with exit flows. An advantage of using PCA is that both entry and exit flows can be analyzed simultaneously, thus allowing us to obtain insight into underlying patterns of

the network-wide flows.

4.4 Results and discussion

4.4.1 Low dimensionality of flows

PCA was performed on the training flow matrix which was specified in the previous section. Since the magnitude of eigenvalues indicates how much variance is explained by the corresponding eigenvector, which is equivalent to PC, a scree plot shown in Fig. 4.3a based on eigenvalues can be leveraged to conduct visual examination. It can be seen through the sharp knee of the curve that the majority of variance contained in the data is virtually contributed by the first few eigen-flows, namely the temporal variability on the first few PCs. Fig. 4.3b further explicitly displays that 8 and 29 PCs, respectively, can account for over 90% and over 95% variance in the data.

There are two possible explanations to such intrinsic low dimensionality of multivariate flow time series. The first one is that it may be attributed to the fact that variation along a small set of dimensions in the original data is dominant. The second reason is that non-negligible correlation among variables may matter greatly, which implies the common underlying patterns or trends across dimensions. In order to understand how each of these two factors accounts for the variance in the current case, PCA can be as well performed on normalized flow variables with zero mean and unit variance. The normalization is specified by equation (4.7). The motivation is that if the low dimensionality still exists after normalization, it can be concluded that the correlation among flows plays the most important role because the normalization procedure has already removed the effect of magnitude in all original flows.

$$\bar{\mathbf{X}}_i = \frac{\mathbf{X}_i - \mu_i}{\sigma_i} \tag{4.7}$$

where μ_i and σ_i denote the sample mean and variance of the i-th column of $\mathbf{X}$.

A comparison between normalized and unnormalized cases is illustrated in Fig. 4.4. No striking difference between two scree plot curves can be seen in Fig. 4.4a, which indicates that the vast majority of low dimensionality of these flows is actually a result of the correlations among them. Moreover, Fig. 4.4b displays that the difference in flow magnitude also accounts for the low dimensionality to some limited extent because more PCs are needed to retain as much variance as before. This important finding coincides with the nature of Shenzhen metro system reflected by Fig. 4.2a in which the CDF curve does not quickly reach a relatively flat level. Such pattern indicates that this is not the case where only a few flows are completely dominant. How these PCs relate to flow and demand patterns are further illuminated in the following section.

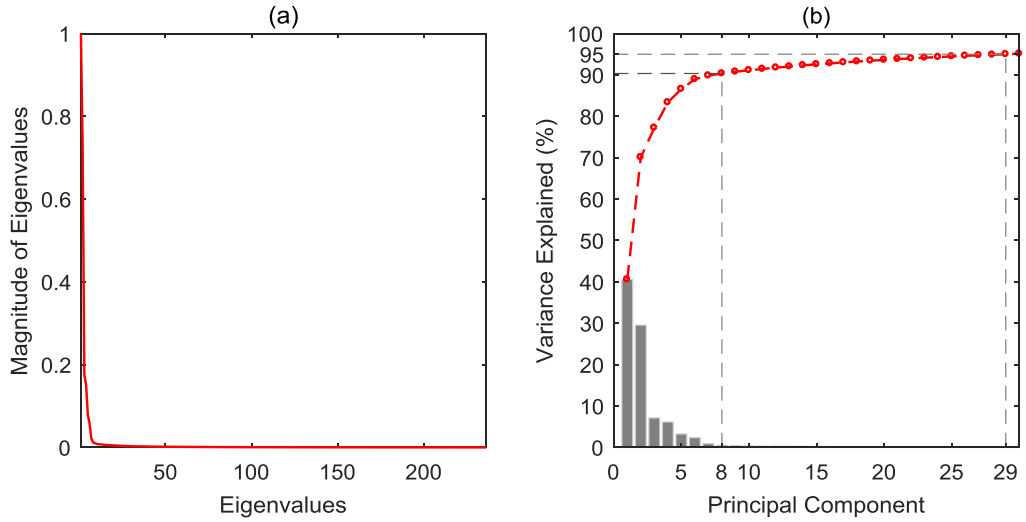


Figure 4.3: Demonstration of the low dimensionality of entry and exit flows. (a) Scree plot of eigenvalues; (b) Cumulative percentage of the total variance explained by PCs (principal components). Over 90% variance can be explained by only 8 PCs, while over 95% can be explained by 29 PCs.

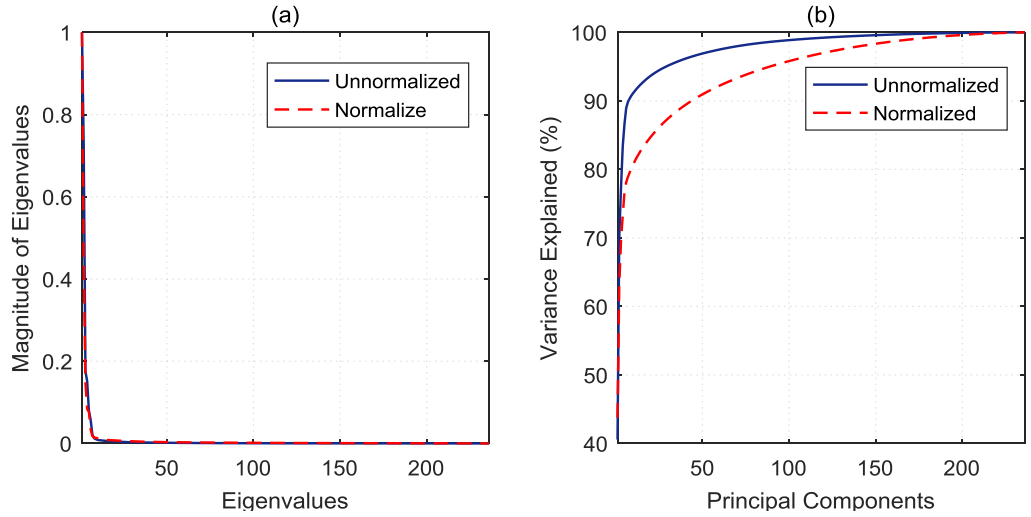


Figure 4.4: Comparison of PCA results for normalized and unnormalized flows. (a) Scree plots based on eigenvalues; (b) Cumulative distribution function plots.

4.4.2 Principal components and eigenflows

Three typical examples of PCs (236 in total) and corresponding eigen-flows are demonstrated in Fig. 4.5. While the top eigen-flow evidently shows weekly periodicity, the other two mostly show randomness with the middle one having two noticeable spikes. Clearly, the first PC very well captures the morning and afternoon peaks of passenger flows in the metro system. The spikes in the middle plot, however, are normally a sign of some special occurrences in the data. Following the taxonomy proposed by Lakhina et al. (2004), the top, middle, and bottom eigen-flows can be roughly referred to as the

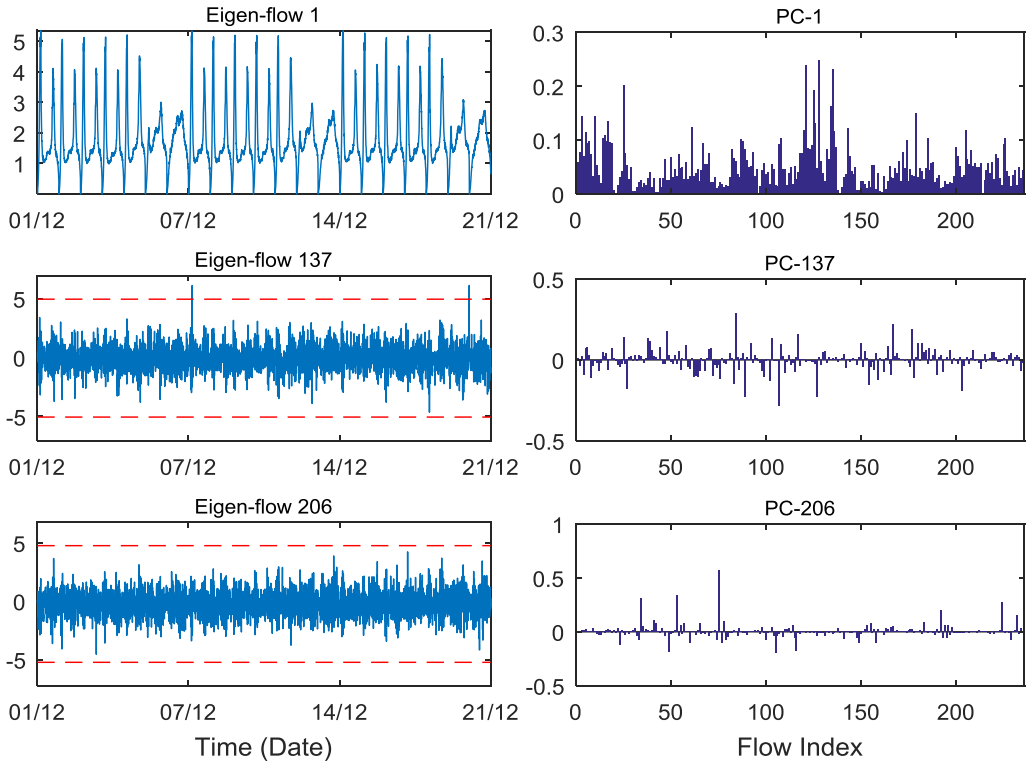


Figure 4.5: Illustration for examples of eigen-flows and PCs.

deterministic, spike, and noise ones, respectively, though in the current case the spike eigen-flows are not always sufficiently significant. This is mainly because there are not many irregular observations of flows in the training set.

The so-called eigen-flows essentially capture all original flows' temporal variation projected onto the PCs. These PCs, shown in the right column of Fig. 4.5, specifically determine how their corresponding eigen-flows contribute to each original flow. Therefore the matrix $\mathbf{V}$ consisting of PCs is also called a *loading* matrix or coefficients. It can be observed that top eigen-flows (variability on top PCs) make greater contribution to original flows. This is consistent with the low dimensionality of original flows.

It can be further investigated how many eigen-flows significantly contribute to one single original flow. This can be done by checking whether a loading coefficient on that row is larger than $\sqrt{p}$. With p equal to 236 in this case, such threshold would be 0.0651. This standard is deemed reasonable because a perfectly equal mixture of all eigen-flows would result in a row of $\mathbf{V}$ with all entries equal, under the condition that columns of $\mathbf{V}$ have unit norm. The result of applying this rule to all rows of $\mathbf{V}$ is illustrated through a CDF plot in Fig. 4.6a. It shows none of the original flows needs more than 70 significant PCs for sufficient reconstruction. In fact, about half of them are composed of less than 45 significant PCs, implying that each entry or exit flow only possesses a relatively small set of temporal variability features.

Furthermore, it is also possible to understand how the magnitude of a single flow is linked to the significant PCs (loading coefficient larger than $\sqrt{p}$) that constitute it. A

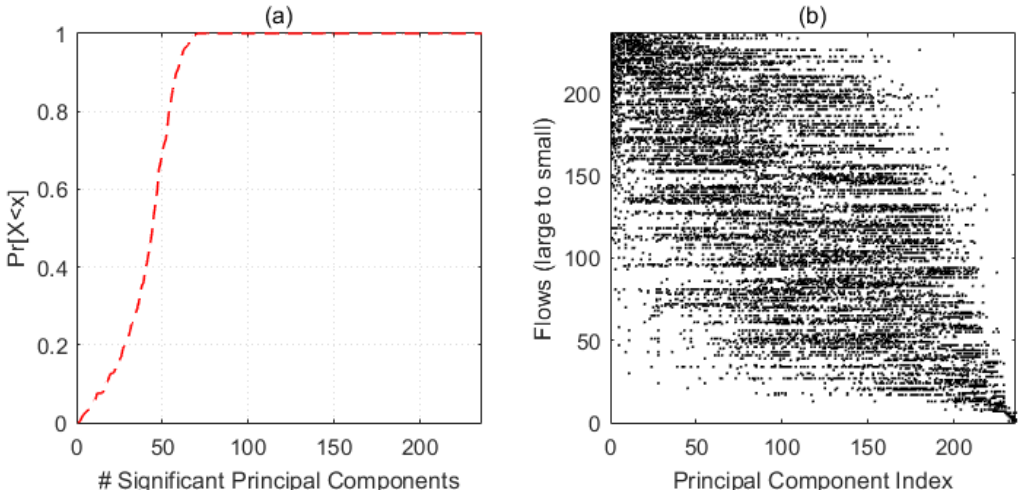


Figure 4.6: Illustration of analysis on flow structure. (a) CDF plot of the number of significant PCs needed for original flows; (b) A scatter plot showing how every single flow is significantly contributed by PCs. The flow index is arranged from top to bottom in a descending order in terms of flow magnitude, while the PC index is arranged from left to right in a descending order in terms of the variance it explains.

graph shown in Fig. 4.6b is used for better visual examination. Its horizontal axis represents the index of PCs ordered in a descending sequence in terms of their eigenvalues (how much variance they explain), and the vertical axis represents the index of original entry and exit flows which are also ordered in a descending sequence in terms of their average magnitude. Once a loading coefficient is larger than $\sqrt{p}$ in absolute value, a black dot will be plotted at that spot. In this manner, the top rows in the graph demonstrate the PCs that are significant in explaining the temporal variability of strong flows, while the bottom rows show those that are significant for weak flows. Although all the dots in Fig. 4.6b scatter considerably, a general diagonal trend from upper left to lower right can be identified. It implies that larger flows tend to be composed of most significant PCs, and vice visa (smaller flows tend to be composed of insignificant PCs). This feature pertains to the approximation of original flows using PCs, which is going to be discussed in the following section.

4.4.3 Reconstructing original flows

According to equation (4.6), the original flows can be approximated using a set of selected PCs. Essentially, such approximation is realized by forming a linear combination of eigen-flows. Fig. 4.7 demonstrates three typical examples of reconstructed flow time series using both 8 (90% total variance explained) and 29 PCs (95% total variance explained). Specifically, the left column displays the overall time series for three weeks, while the right one zooms into a more detailed level with only one day illustrated.

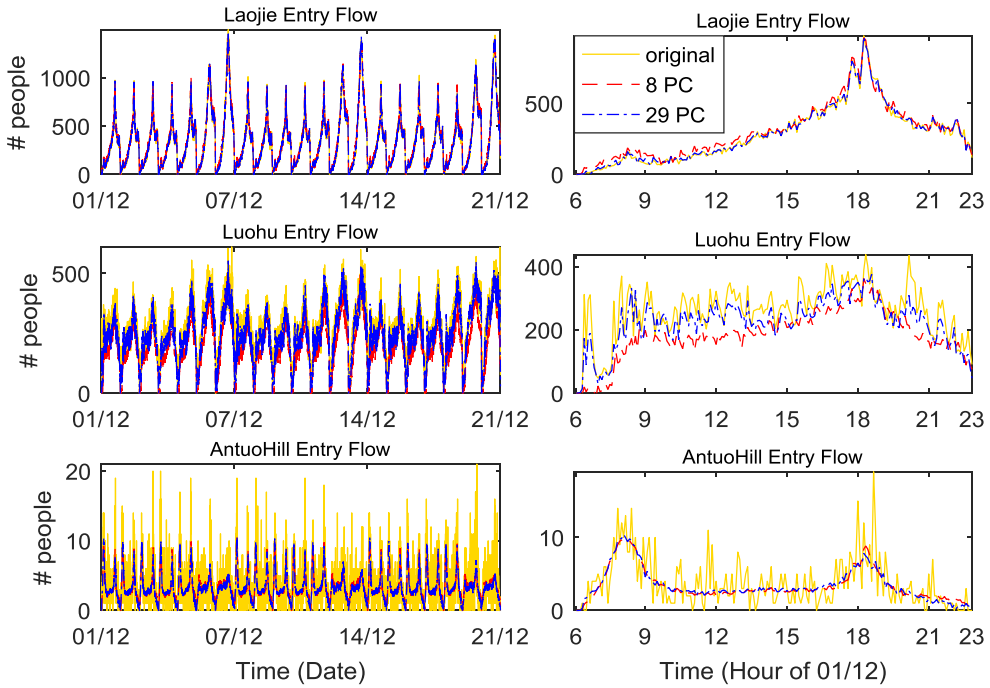


Figure 4.7: Examples of approximating original flows using different number of PCs. The left column illustrates the results of the entire period covered by the training data, while the right column shows the zoom-in plots of the first day (December 1, 2014).

As the busiest metro station in Shenzhen, Laojie station's entry flow profile (top) can be well reconstructed in both scenarios. The middle row, however, shows a different case where the original data (Luohu entry flow) can be approximated much better with 29 PCs than 8 PCs. Note Luohu is a metro station heavily used by travelers because of the train station and port connecting Shenzhen and Hong Kong next to it. The bottom row then shows something different from both above. It can be seen that the entry flow at AntuoHill station is very low all the time, and both approximations cannot capture detailed fluctuations but the main trend of the time series. These results are in line with the low dimensionality of flows and the implication obtained from the last section that larger flows tend to be composed of most significant PCs, while smaller ones tend to be composed of insignificant PCs.

4.4.4 Temporal stability of flow structure

With the whole data set divided into training and validation parts, it can be examined whether the PCs derived from the previous time period can be also used to approximate future time series. It is valuable to know if the underlying flow patterns are stable over time so that this intrinsic feature can benefit multiple applications, such as anomaly detection and short-term prediction. In practice, the PCs contained in the loading matrix $\mathbf{V}$ was computed using the training set, and then these PCs were used to derive eigen-flows of the validation data set based on equation (4.4). Eventually, the same

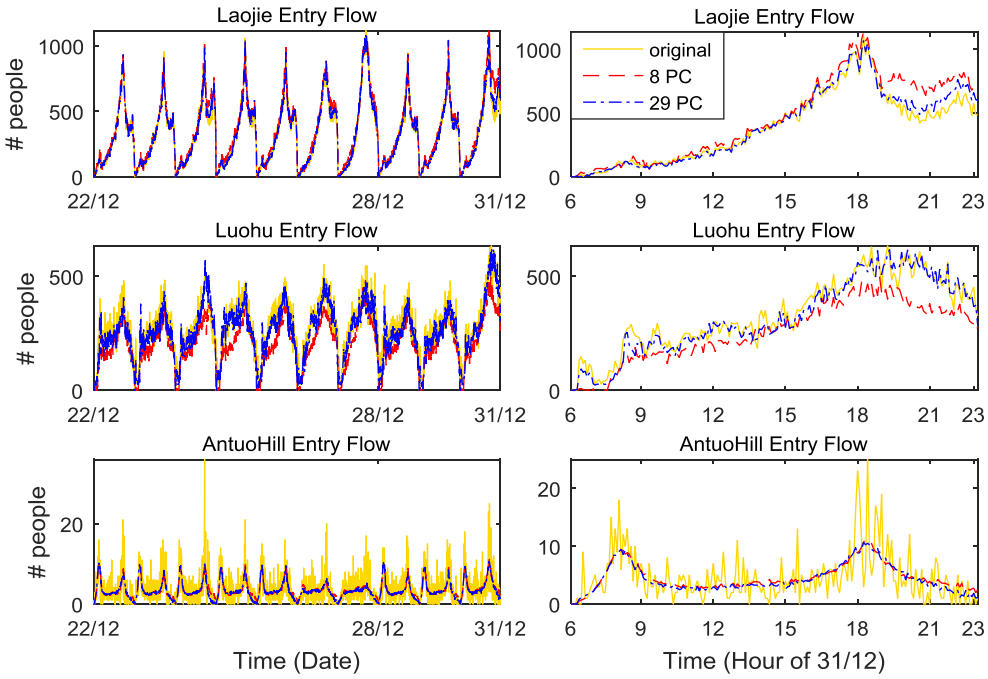


Figure 4.8: Examples of approximating flows using PCs that are not computed based on these flow data. The left column illustrates the results of the entire period covered by the validation data, while the right column shows the zoom-in plots of the last day (December 31, 2014).

approximation procedure can be performed using equation (4.6), of which results are shown in Fig. 4.8.

Overall, it can be observed that the PCs computed using past data are still capable of approximating the profile of future data. Nevertheless, it should be noticed that the effectiveness of these "old" PCs can be diminished when non-recurrent or special events occur, as shown in the top row of Fig. 4.8 which demonstrates the entry flows of Laojie station on the new year eve (December 31, 2014). With more incoming passengers than an average weekday, it can be seen that the approximation results of both cases (8 and 29 PCs) illustrate more deviation from the original flow as well. The comparison between the top right plots of Fig. 4.7 and Fig. 4.8 makes it even clearer. As to the other two examples, however, the distinction is not as clear as the case of Laojie station. It can thus be concluded that the structure of flows captured by PCs can still be temporally stable, mostly for normal days, but failure can occur when there are uncommon changes in flows.

4.5 Conclusion

This chapter demonstrates how PCA can be applied to multivariate PT passenger flows as a solution to high-dimensional data analysis problems. With a one-month AFC data

from the Shenzhen metro system leveraged, PCA is performed on a 4284×236 multivariate time series matrix (entry and exit flows considered simultaneously) transformed from the original individual tapping records. The results and analysis show that a great amount of variance contained in the original data can be effectively retained in lower-dimensional sub-spaces composed of top few PCs. This feature of low dimensionality is thus thoroughly examined in the study, with the essence of PCs and eigen-flows, as well as the temporal stability of PCs revealed in the subsequent investigation.

The following directions can be explored by future research. First, it is worthwhile examining how PCA can be integrated into modeling and prediction of large-scale passenger flows in PT systems. Reducing the high dimensionality cannot be bypassed for this type of research. Second, some other dimensionality reduction techniques, particularly those nonlinear ones, could also be tested to see whether they can outperform PCA in this application. Third, this study is limited to treating the temporal aspect as only one dimension. In fact, the concept of tensor can be incorporated to separate the temporal dimension into two, including within-day and across-day dimensions. Tensor decomposition techniques can then be applied to extract those principal patterns across different temporal patterns. This could potentially be more applicable for large-scale modeling and predictions.

Chapter 5

Clustering of Public Transport Stops

In the previous chapter, we have demonstrated that PCA can serve as an effective technique to address the issue of high dimensionality for studying large-scale passenger flows in PT networks. In this chapter, we offer another solution from a different perspective, taking into account the network characteristics. To this end, we propose a *k*-means-based method to cluster PT stops for constructing zone-to-zone OD matrices. The key is to heuristically determining the number of clusters by combining spatial distances and passenger OD flows. Differing from the traditional way of grouping stops based on predefined traffic analysis zones, our proposed method provides a data-driven perspective for solving such problems using passenger flows that can be directly observed rather than their proxies. The method is demonstrated by a case study of the PT system of The Hague, the Netherlands. This clustering approach is particularly suitable for urban areas with high density of PT stops, because it allows for integrating travelers' (origin and destination) stop and route choices into subsequent modeling studies.

This chapter is outlined as follows. Section 5.1 introduces the research background and objectives with a review of relevant literature. Section 5.2 describes how passenger journeys were constructed for the data of The Hague. We then detail the proposed methodology in section 5.3, followed by the presentation of the results in section 5.4. Section 5.5 finalizes this chapter with the conclusion and suggestions for future research.

This chapter is an edited version of the following article:

Luo, D., Cats, O. & van Lint, H. (2017) Constructing transit origin-destination matrices with spatial clustering. *Transportation Research Record*, 2652(1), 39-49.

5.1 Introduction

Demand analysis and modeling is an significant component for PT planning. The purpose of such research is to estimate and evaluate passenger demand by using models and by collecting and analyzing data pertaining to current and future PT needs (Ceder, 2015). Traditionally, the classical sequential four-step process has been extensively used in both academia and practice to estimate the aggregated travel demand for a number of traffic analysis zones (TAZ) which are predetermined based on geographical and socio-economic factors. Following this process, the share of PT demand is then computed at the step of modal split or mode choice using discrete choice models. This four-step method provides researchers and practitioners with a straightforward way to obtain PT demand when such demand could hardly be observed directly, though the results cannot always be desirably accurate. However, with AFC systems being adopted by more and more PT agencies, a new type of data source is rapidly becoming available. AFC systems record individual travelers' boarding and/or alighting information, greatly facilitating the research on passenger travel patterns that can support PT network planning, behavioral analysis and PT demand estimation and forecasting (Pelletier et al., 2011).

A large amount of research effort has been directed to PT OD estimation, especially for the cases where only entry or exit information is available, or even neither of them is available. Different methodologies have been proposed to infer OD matrices for PT journeys with limited boarding or alighting information (Barry et al., 2002; Trépanier et al., 2007; Alfred Chu & Chapleau, 2008; Nassir et al., 2011; Wang et al., 2011; Munizaga & Palma, 2012; Gordon et al., 2013; Ma et al., 2013), and these methods can be categorized based on their estimation assumptions, including walking distances (buffer zones), transfer times, and last destination assumptions (Alsger et al., 2016). Along with the increase in the number of methods, the importance of evaluation and validation on the OD estimation methods and results has also been highlighted in a series of studies (Farzin, 2008; Barry et al., 2009; Devillaine et al., 2012; Munizaga et al., 2014; Alsger et al., 2015). Recently, Alsger et al. (2016) used a high-quality dataset containing accurate boarding and alighting information to validate a multi-leg journey inference method, where the alighting information is assumed to be unknown for validation sake.

The aforementioned OD estimation studies focused on attaining more accurate journey inference, whereas less effort has been directed toward stop or station aggregation while constructing the demand matrices. Stop or station aggregation, in this context, means that PT users' activities of originating from or destining to an individual stop can be virtually associated with an area which covers a number of adjacent PT stops or stations (Lee et al., 2013). In this sense, the demand at a more aggregate level can be of more practical use for both PT researchers and practitioners. McCord et al. (2012) pointed out that the size of stop-to-stop OD matrices makes it difficult to synthesize important flow patterns and to estimate stop-to-stop OD passenger flows accurately. By

grouping PT stops, however, the estimation, analysis, and communication of passenger flows can be improved. Furthermore, understanding the PT demand at an aggregate level was also motivated by Lee et al. (2013). They highlighted that the ability to define a specific land-use type and the temporal characteristics related to passengers' activities can be enhanced through the stop aggregation. In addition, the aggregation of stops is also in line with the analysis and modeling of PT users' stop choice behavior which has been explored by several recent studies (Hassan et al., 2016; Nassir et al., 2015). The rationale is that in reality, travelers are very often capable of choosing from a set of origin and destination stations which are within their acceptable distance. As a result of such behavior, different choices in PT services (modes and routes) can be characterized in terms of travel demand from one area to another.

A limited number of studies involving the stop aggregation can be identified in the current literature. For instance, Chu & Chapleau (2010) used the term "anchor points" to define the places that a person repeatedly visits, which usually include residence, work, or study locations. They performed spatial aggregation by grouping stops within 50 m of each other to form a new node. A so-called "stop-aggregation model" was later proposed and applied to studying the metro PT of the Minneapolis-Saint Paul metropolitan area in Minnesota (Lee et al., 2012; Lee & Hickman, 2013). This model aims to define a generalized definition of a "stop" that more closely matches the nature of locations serving as passenger origins and destinations. An aggregated area around a PT stop or station can thus be defined by three parameters: (a) distance or proximity, measured by using Euclidean and network distances in geographic information systems; (b) text in the description of the stop, queried using database tools in SQL; (c) the catchment area which is defined as how a stop relates to the land uses surrounding it. Alsger et al. (12) simply aggregated the estimated OD trips according to the 1,515 zones in the Brisbane Strategic Transport Model to provide an overview of the results. McCord et al. (2012) proposed two computationally efficient heuristic algorithms to aggregate bus stops at the route level in order to reduce the size of the OD matrix for improved estimation, analysis, and communication. More recently, Tamblay et al. (2016) developed a methodology to estimate zonal OD matrix for a PT system. Based on a stop-to-stop OD matrix created by using smart card data from Santiago de Chile, a Logit model was constructed to compute the probability that an observed trip using PT stops k and l (as their boarding and alighting points, respectively), was originated in zone i and ended in zone j. It is notable that in their methodology a zonal system must be pre-defined and a survey is required to help identify the model parameters.

Unsupervised learning techniques have recently been employed to investigate spatial travel pattern and demand given their natural advantages in solving clustering problems. One of the successful applications turns out to be the identification of individual PT riders' spatial and temporal travel patterns using the density-based spatial clustering of applications with noise (DBSCAN) algorithm (Ma et al., 2013; Kieu et al., 2015a,b). This specific algorithm stands out in its flexibility. It does not require predetermining the number of clusters, and can identify arbitrarily shaped clusters. Ma et al.

(2013) first applied this algorithm to examine the spatial travel pattern of PT users in Beijing after inferring individuals' journey chains. Based on this, Kieu et al. (2015b) performed a two-level approach based on the standard DBSCAN algorithm to reveal both spatial and temporal travel patterns in South East Queensland, Australia. They further improved the efficiency of this approach by utilizing the existing knowledge of individual travel pattern while clustering the studied journey, developing a so-called Weighted-Stop DBSCAN (WS-DBSCAN) algorithm Kieu et al. (2015a).

The abovementioned studies highlight the importance of stop or station aggregation in analyzing AFC data. In many cases where PT services are well provided in both urban and suburban areas in terms of number of stops and routes (e.g., The Hague, as shown in Figure 2.2), the information of traveler OD flow from one area to another, both of which contain a group of bus or tram stations, would be much better usable in PT modeling, prediction and management than by examining them at individual stop level. This paper proposes a *k*-means-based station aggregation methodology which can in four steps quantitatively determine the clustering by considering both flow and spatial distance information. This method is applied to a case study of The Hague, the Netherlands, by specifying a criterion that the ratio of average intra-cluster flow to average inter-cluster flow should be maximized while maintaining the spatial compactness of all clusters. In the first step, we obtain a number of different clustering scenarios by implementing the standard *k*-means algorithm with the geodesic distance considered as the only feature. Then, two metrics based on spatial distance and passenger flow respectively are computed, and finally integrated to determine the optimal number of clusters. The proposed data-driven method allows us to obtain clusters that are on one hand sufficiently large to enable the consideration and modeling of travel alternatives between parts of the network, while on the other hand being compact enough to include only viable alternatives and support fine-grained demand estimation. Unlike the standard DBSCAN algorithm that can identify some points as "noise" which are then not included in any of the clusters, the proposed method assures that all travel information is retained in the OD matrix attained.

5.2 Constructing passenger journeys

In this chapter, we use the AFC data from The Hague introduced in section 2.2.2. The first task was to construct multi-ride journeys, which was performed by Bagherian et al. (2016). The procedure started with excluding data that contains missing values (e.g. missing tap-in or tap-out, no line identifier). Three types of transactions were subsequently removed, including those with identical location of tap-in and tap-out; those between stops i, j with $t^o_{n,i} - t^c_{n,j} < \gamma^{leg}_{min}$, where $t^o_{n,i}$ and $t^c_{n,j}$ represent the tap-in and tap-out time of a card ID n at station i and j, respectively. γ^{leg}_{min} denotes the minimum duration of a leg; those with abnormally long duration $t^o_{n,i} - t^c_{n,j} > \gamma^{leg}_{max}$, where γ^{leg}_{max} denotes the maximum duration. After this step, transactions were grouped using card

ID for each day in the analysis period, and within each group the transactions were also sorted using check-in timestamp. Finally, an iterative procedure chained transactions forming a journey if $t^c_{r_n} - t^c_{r-1_n} < \gamma^{transfer}$, here $t^c_{r_n}$ and $t^c_{r-1_n}$, respectively, represent the tap-in time of transaction r and the tap-out time of transaction $r-1$. $\gamma^{transfer}$ denotes the time interval between two successive legs with same card ID. In this case study, γ^{leg}_{min}, γ^{leg}_{max}, and $\gamma^{transfer}$ were set to be 1, 60 and 35 min, respectively. Once a journey was formed, transfer times and the number of transfers were also computed, and the journey was added to the database. Consequently, more than six million journeys in the analysis period were generated. The output of this procedure is a database of the identified journeys, including an ID, date, number of transfers, and a list of details (route id, tap-in time and location, tap-out time and location) for all the rides.

An issue regarding the validity of journey identification was observed while scrutinizing through the dataset as some journeys are unreasonably long (i.e. several hours). These "noise" inferred multi-leg journeys are presumably caused by short activity chaining and were thus removed from the dataset by adopting a threshold of maximum 90 minutes. This value was determined based on the maximum time that a person needs to spend reaching his/her destination in Haaglanden. As a result of this cleaning process, 14,794 journeys were removed, leaving 99.76% of the original records for further analysis.

5.3 Methodology

5.3.1 A four-step *k*-means-based method

The *k*-means algorithm has been extensively applied in various fields since McQuenn (1967) proposed it. Given a set of n observations $(\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n})$, each of which is a d-dimensional real vector, this clustering algorithm aims to partition the n observations into K ($\leqslant n$) mutually exclusive and collectively exhaustive clusters $\mathbf{C} = \{C_1, C_2, ..., C_K\}$. It iteratively determines the center $\boldsymbol{\mu}_i$ for each cluster C_i and assigns each observation to a cluster whose center is closest to the observation. This iterative clustering process terminates when the assignments no longer change, which can be described as to minimize the within-cluster sum of squares (sum of distance functions of each observation in the cluster C_i to the center $\boldsymbol{\mu}_i$):

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} ||\mathbf{x} - \boldsymbol{\mu}_i||^2 \tag{5.1}$$

For details about the implementation of the *k*-means algorithm, readers are referred to Tou & Gonzalez (1974). Its main disadvantage is that the number of clusters, K, must be supplied as a parameter. In this study, a four-step *k*-means-based station aggregation method is proposed, in which a quantitative way to determine the optimal K is incorporated as described in Figure 5.1.

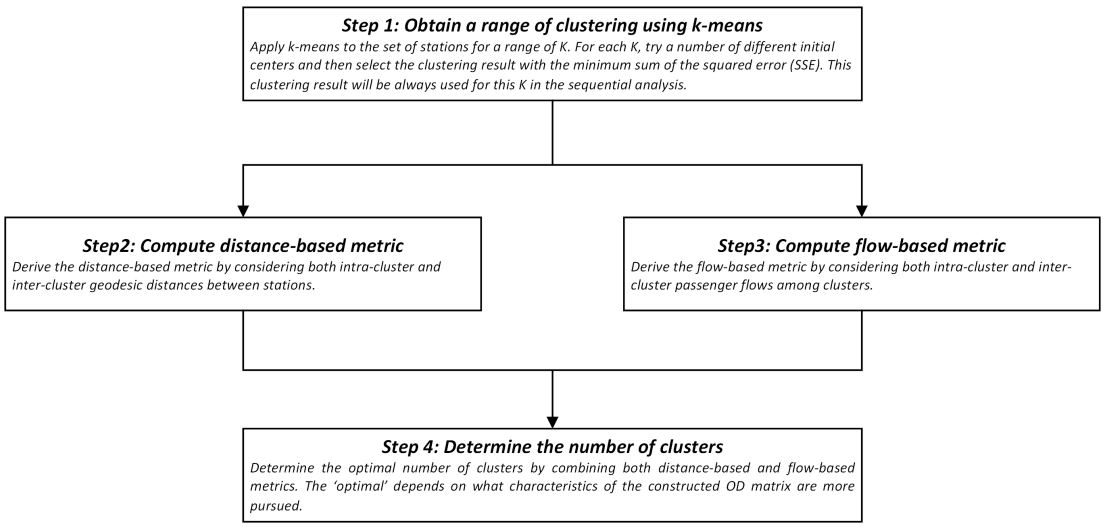


Figure 5.1: Illustration of the proposed *k*-means-based stop aggregation method.

The method starts with finding the best clustering based on a chosen measure for each K, and then continues with the computation of two metrics related to spatial distance and passenger flow respectively. In the final step, two metrics are integrated for the determination of the optimal number of clusters K^*. Such a method is flexible as it can accommodate different formulations of both metrics and final integration function in order to cater different purposes pertaining to the construction of PT OD matrix. The essential idea, however, is to maximize either the intra-cluster or the inter-cluster flow while maintaining the spatial compactness of all clusters simultaneously.

5.3.2 *k*-means-based clustering

Considering that the clusters of PT stations should be spatially compact, the geodesic distance between points, which can be calculated based on their coordinates, is used as the only feature in the *k*-means clustering. While implementing the *k*-means algorithm, a set of K points are input as the initial cluster centers so that the algorithm can proceed with iterations itself. Since the result of the *k*-means algorithm can vary given different initial centers, a common way to obtain better and reproducible results is to perform the algorithm a number of times with different initial centers and select the initial centers which produces the optimal clustering in terms of the adopted measure. In this study, a measure called sum of the squared error (SSE) is employed to help select the initial centers because it can reflect the quality of a clustering. The lower SSE is, the better the clustering. The SSE is defined as follows.

$$SSE(K) = \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} d^2_{\mu_i, x} \tag{5.2}$$

where $d_{\mu_i,x}$ denotes the geodesic distance between a station and the cluster center to which it belongs. The *k*-means algorithm was programmed in Python and the imple-

mentation process is described as follow: A number of randomly generated sets of initial centers were tested for each K ranging from 2 to 30 in the current study, and the initial center scenario which resulted in the minimum SSE was eventually selected and fixed for this K in the sequential analysis. Given the particular spatial distribution of stations in the study area (i.e., most stations are in the core area of The Hague with other scattered in relatively isolated areas), six subareas, including Delft, Zoetermeer, northeast, northwest, southwest, southeast of The Hague were set up. When the number of cluster K was larger than 5, two initial center points were randomly generated from the Delft and Zoetermeer subareas and the rest would also be generated from The Hague subareas. By doing so, the efficiency of implementing the k-means algorithm for a great number of iterations was dramatically improved in this case study. However, it is still worth mentioning that it can be time-consuming to complete all clustering experiments for a large number of K (>25). This issue can be further resolved by optimizing the k-means program.

5.3.3 Distance-based metric

The construction of the distance-based metric adopts the approach proposed by Ray & Turi (1999). It examines the spatial compactness of a clustering by taking into consideration both intra-cluster and inter-cluster distance measures. The former one computes the square of distance between a point and its cluster center, and then takes the average of all of them, denoted by D^{intra}:

$$D^{intra} = \frac{1}{N} \sum_{i=1}^{K} \sum_{\mathbf{x} \in C_i} d^2_{\mu_i, x} \tag{5.3}$$

where N is the number of stops in the network.

The inter-cluster distance measure, D^{inter}, on the other hand, only takes the square of minimum distance between cluster centers because as long as the minimum of such distance is maximized, the others will by definition be larger than it. This measure is defined as follows:

$$D^{inter} = \min d^2_{\mu_i, \mu_j}, \forall i \neq j \tag{5.4}$$

The two measures are then combined by taking the ratio as follows:

$$\tau = \frac{D^{intra}}{D^{inter}} \tag{5.5}$$

where τ denotes the final distance-based metric. To obtain the optimal number of clusters in terms of spatial compactness, τ is minimized since the intra-cluster distance measure D^{intra} in the numerator should be minimized while the inter-cluster distance measure D^{inter} in the denominator should be maximized.

5.3.4 Flow-based metric

The passenger flow at the stop level can be first derived from the original dataset and then be aggregated based on a specific clustering. The flow-based metric provides additional information that can be utilized to determine the optimal number of clusters. Intuitively, total intra-cluster flow decreases as the number of clusters grows given the constant total flow over the entire study period. More flow are naturally assigned to the inter-cluster one (see Figure 5.2b).

When considering the flow information, we can either seek to maximize the total inter-cluster flow over total intra-cluster one or vice-versa depending on the application and the analysis objectives. An argument in favor of the former case is that it leads to more flow being assigned as inter-cluster (non-diagonal) elements in the OD matrix. In contrast, by making the intra-cluster flow more significant, most self-contained and coherent clusters in terms of travel demand (diagonal elements) can be obtained, which is more desirable from a planning perspective. In the current case of The Hague, the second option was eventually adopted and the following two flow measures are proposed:

$$F^{intra} = \frac{1}{K} \sum_{i=1}^{K} \sum_{x_m, x_n \in C_i} f_{x_m, x_n} \tag{5.6}$$

$$F^{inter} = \frac{1}{K^2 - K} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{x_m \in C_i, x_n \in C_j, \forall i \neq j} f_{x_m, x_n} \tag{5.7}$$

where f_{x_m,x_n} denotes the passenger flow from stop x_m to x_n and K denotes the number of clusters. F^{intra} and F^{inter}, essentially, represent the average intra-cluster and average inter-cluster flow, respectively (see Figure 4c). To combine two measures, the ratio of F^{intra} and F^{inter} is adopted and defined as follows:

$$\delta = \frac{F^{intra}}{F^{inter}} \tag{5.8}$$

where δ denotes the flow-based metric. To obtain most self-contained clusters, δ should be maximized so that the average intra-cluster flow is as significant as possible.

5.3.5 Determining the number of clusters

To determine the optimal number of cluster with both distance-based and flow-based metrics, different objective functions can be formulated. Since in the current case we aim to (i) obtain clusters that are as spatially compact as possible, which can be achieved by minimizing τ; (ii) attain an intra-cluster flow as strong as possible, which can be achieved by maximizing δ, a straightforward way that takes the ratio of δ to τ

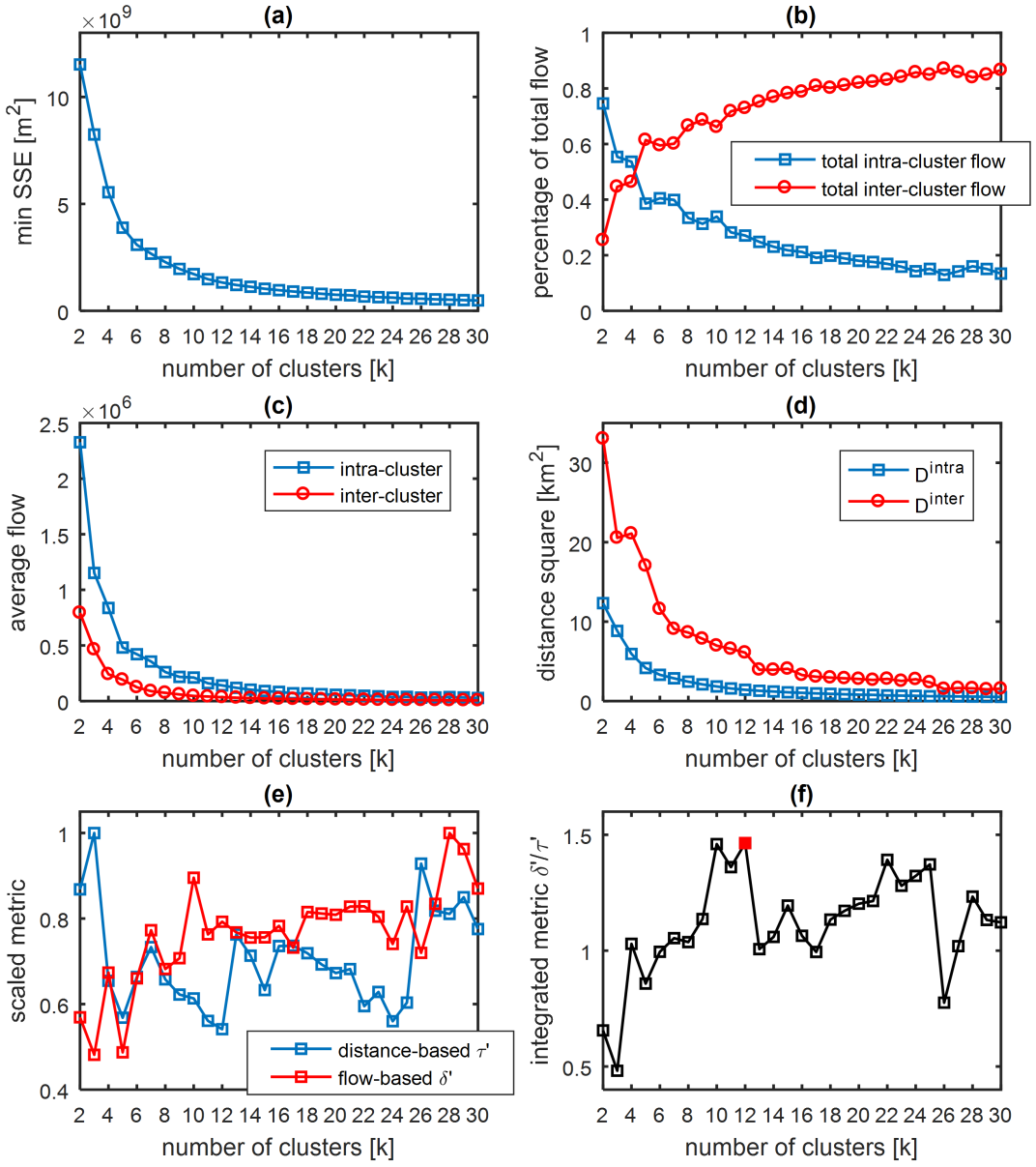


Figure 5.2: (a) SSE decreases exponentially as the number of clusters increases (SSE = sum of the squared error); (b) percentage variation in both total intra-cluster and total inter-cluster flows; (c) variation in both average intra-cluster and average inter-cluster flows; (d) illustration of intra-cluster and inter-cluster flow measures; (e) illustration of two scaled metrics; (f) the integrated metric which reaches the maximum value when the number of cluster is equal to 12.

is adopted. A scaling procedure is applied to both metrics before taking the ratio so that their magnitudes are comparable.

$$X' = \frac{X}{X_{max}} \tag{5.9}$$

After applying the scaling procedure, the optimal number of clusters K^* is attained:

$$\underset{K\in[K_{min},K_{max}]}{\operatorname{argmax}} \frac{\delta'_K}{\tau'_K} \tag{5.10}$$

where δ'_K and τ'_K denote the scaled flow-based and distance-based metrics for the K clustering, respectively.

5.4 Results and analysis

5.4.1 Results

Figure 5.3 shows the clustering results determined for each K ranging from 2 to 30 in this study based on the calculation of SSE. The different clusters are illustrated using various combinations of colors and markers without the underlying PT network included. The variation in SSE is presented in Figure 5.2a. It can be seen that as the number of clusters increases, more clusters are generated mainly within The Hague area. SSE does not decrease linearly as K increases. Instead, a sharp drop can be observed in 5.2a when K is approaching 8 and then the decline becomes increasingly flat as K grows.

Figure 5.2d reveals that both the intra-cluster and inter-cluster distance measures show a decrease pattern as K grows, although the intra-cluster one is smoother than its counterpart. Two scaled metrics are plotted together in Figure 5.2e for the sake of comparison. No specific patterns are very clear for both metrics, but when K is equal to 5, 12, 24, the distance-based metric reaches some local minimums. The flow-based metric exhibits an overall growing pattern.

The integrated metric which takes the ratio of scaled flow-based to scaled distance-based metric is displayed in Figure 5.2f. The optimal number of cluster in terms of the integrated index in this case turns out to be 12 (highlighted in Figure 5.2f) albeit there is only a very slight difference between $K10$ and $K12$, and $K22$, 23, 24 and 25 are also close. Detailed results and analysis of this optimal clustering are presented in Figure 5.4, including the spatial outcome and the number of stations contained in each cluster. The bar chart shows that the number of stations contained in the more isolated parts of the network (Clusters 1,2 and 7) is significantly lower than other parts of the network. This is arguably attributed to the low density of stations in these areas. Within the core area of The Hague, stations are more evenly distributed in different clusters, although there are still more stations assigned in Cluster 5 and 8.

Aggregated passenger demand at the cluster level are shown in Figure 5.4d and Figure 5.4e, with the former specifying all the numbers while the latter performing the visualization through a chord chart. Apparently, Cluster 5 accounts for the most demand because it contains all stations around the central station of The Hague with connections to train services. It is followed by Cluster 11 and 12, the former of which covers

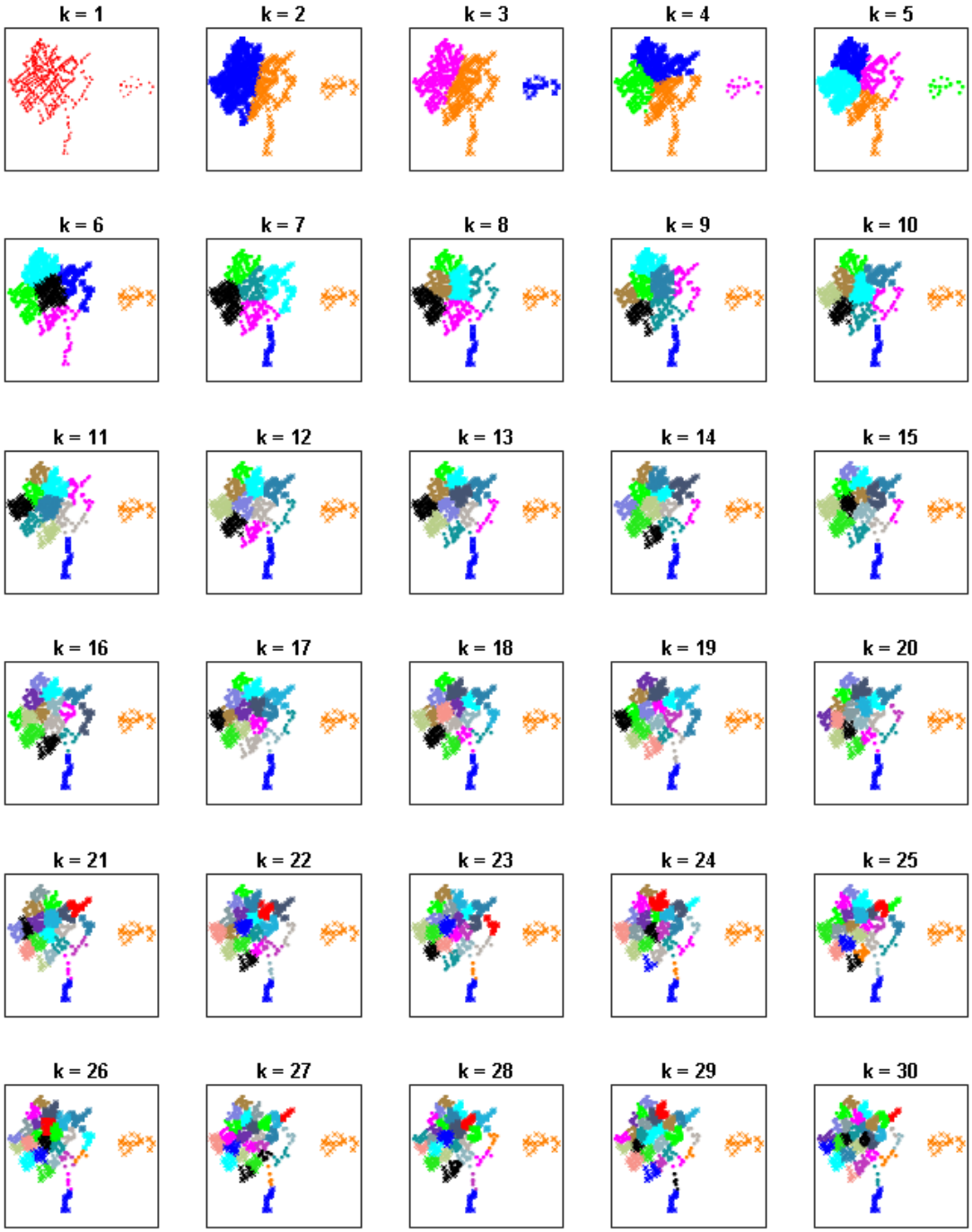


Figure 5.3: Illustration of clustering results with different K.

the area of Den Haag Holland Spoor which is the second biggest train station in The Hague, while the second of which covers the main commercial area. Cluster 1, 3, 4, 7 and 10, however, show relatively low demand for HTM services, which can be partially explained by competing PT operators (i.e. bus and train). Furthermore, the low demand of Cluster 4 and 10 is attributed to the relatively lower-density residential areas and lower overall PT market share. Cluster 8 exhibits a higher demand, presumably due to the presence of regional and national institutions (e.g. museums, theater,

stadium and embassies) which attract visitors.

5.4.2 Spatial variability analysis

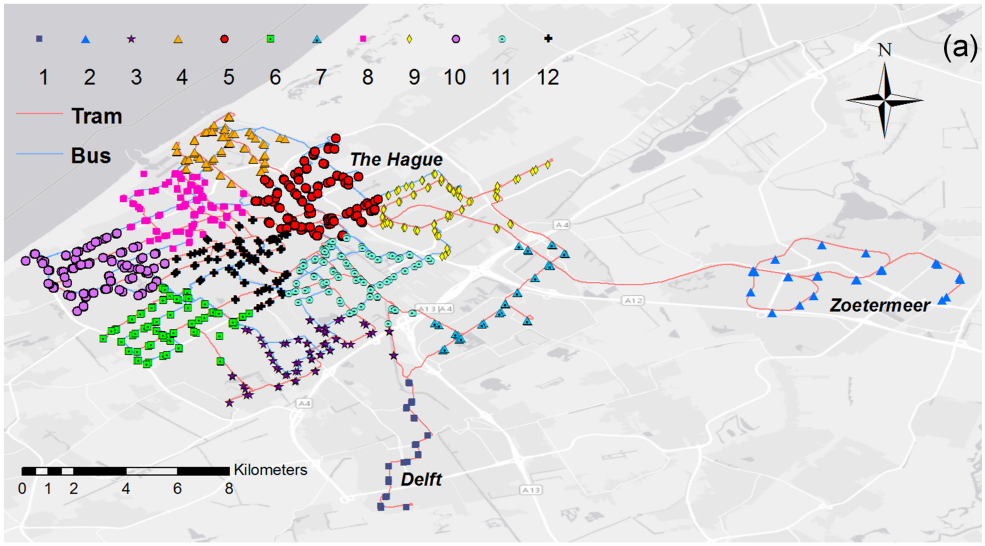
The spatial variability of all individual clusters is investigated and illustrated in Figure 5.4c through boxplots of the spatial distance between stations in a cluster. This is important in a sense that travelers are assumed to be able to reach alternative stations as easily as possible within a cluster using non-motorized modes, primarily walking and cycling. If the spatial variability of a cluster is too large, then it means some stations within the cluster are too far from each other and would not be likely be considered as alternatives by travelers. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the + symbol. All clusters' median station-to-station distances are below 2 km, but Cluster 1, 2, 7 particularly show larger variability due to their larger spatial extents. Besides these three spatially isolated clusters, Cluster 3 and 9 also show more variability than the average. This is because these two clusters are generated with more scattered stations and they do not show the most desired round shapes as a result of the *k*-means algorithm. For example, some stations in the west of Cluster 9, as well as some in the southwest of Cluster 3 are more distant from the majority of the stations. This is admittedly one of the drawbacks of adopting the *k*-means algorithm.

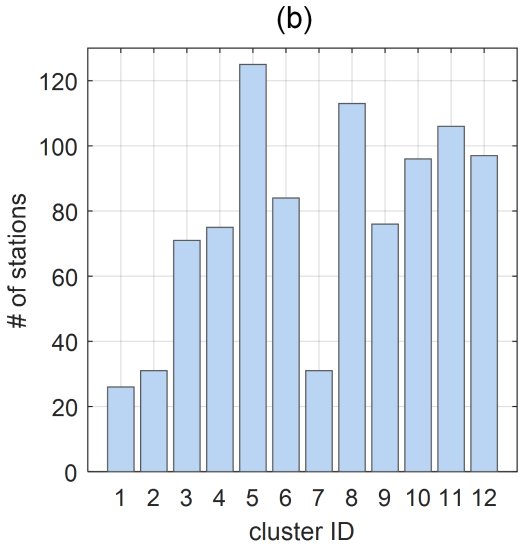
5.4.3 Temporal variability analysis

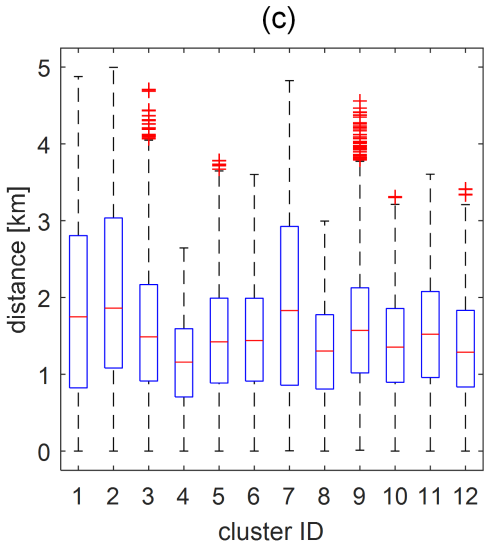
As Figure 5.5a shows, the PT demand in The Hague reveals clear within-day and across-day patterns during the regular operating time. To examine the influence of time-dependent passenger flow on the determination of number of clusters, the proposed method was also implemented with multiple temporal passenger flow profiles from different periods. The following periods were investigated:

Weekdays

- Morning prepeak, before 7:30 a.m.;
- Morning peak, 7:30 a.m. to 10:00 a.m.;
- Midday, 10:00 a.m. to 3:00 p.m.;
- Afternoon peak, 3:00 p.m. to 7:30 p.m.;
- Afternoon postpeak, after 7:30 p.m.







(d)

OD	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10	Cluster 11	Cluster 12
Cluster 1	121177	0	1546	5850	23001	0	13751	228	2711	0	22223	0
Cluster 2	0	178816	0	0	66471	3935	40878	3897	35243	2419	0	15844
Cluster 3	1069	0	58100	1232	35804	16170	0	4360	4071	4873	45274	1661
Cluster 4	5966	0	1187	28283	117664	4957	0	23469	7608	4253	28012	5818
Cluster 5	30268	87636	44599	141126	466499	100297	70813	153711	202626	96113	228770	272318
Cluster 6	0	4668	25186	6124	80867	122933	3605	7819	7006	32956	62252	110581
Cluster 7	10912	42561	0	0	48622	2801	42323	2783	23480	1794	18607	9671
Cluster 8	212	4435	4631	26360	132465	7182	3003	49039	7032	36745	47954	53815
Cluster 9	2279	34005	3808	7319	150812	6457	23575	8935	105280	9430	6548	46555
Cluster 10	0	2551	5646	4222	79136	19509	1850	28635	7529	48343	10960	48864
Cluster 11	40727	0	55729	36417	206610	62525	25834	52833	7128	7718	196911	126246
Cluster 12	0	20964	1856	9366	198927	115379	13404	57849	51066	45848	99807	272591

(e)

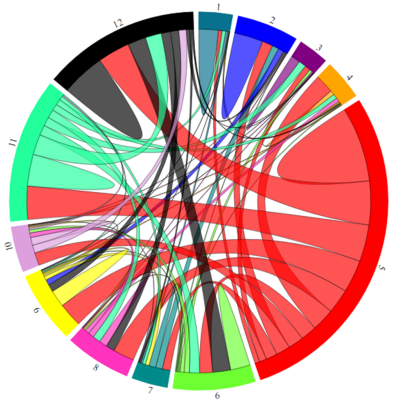


Figure 5.4: Illustrations of the optimal clustering (K=12) (a) visualization of 12 clusters; (b) number of stations contained in each cluster; (c) illustrations of clusters' spatial variability (d) resulting OD matrices over the entire study period; (e) visualization of the OD passenger flow.

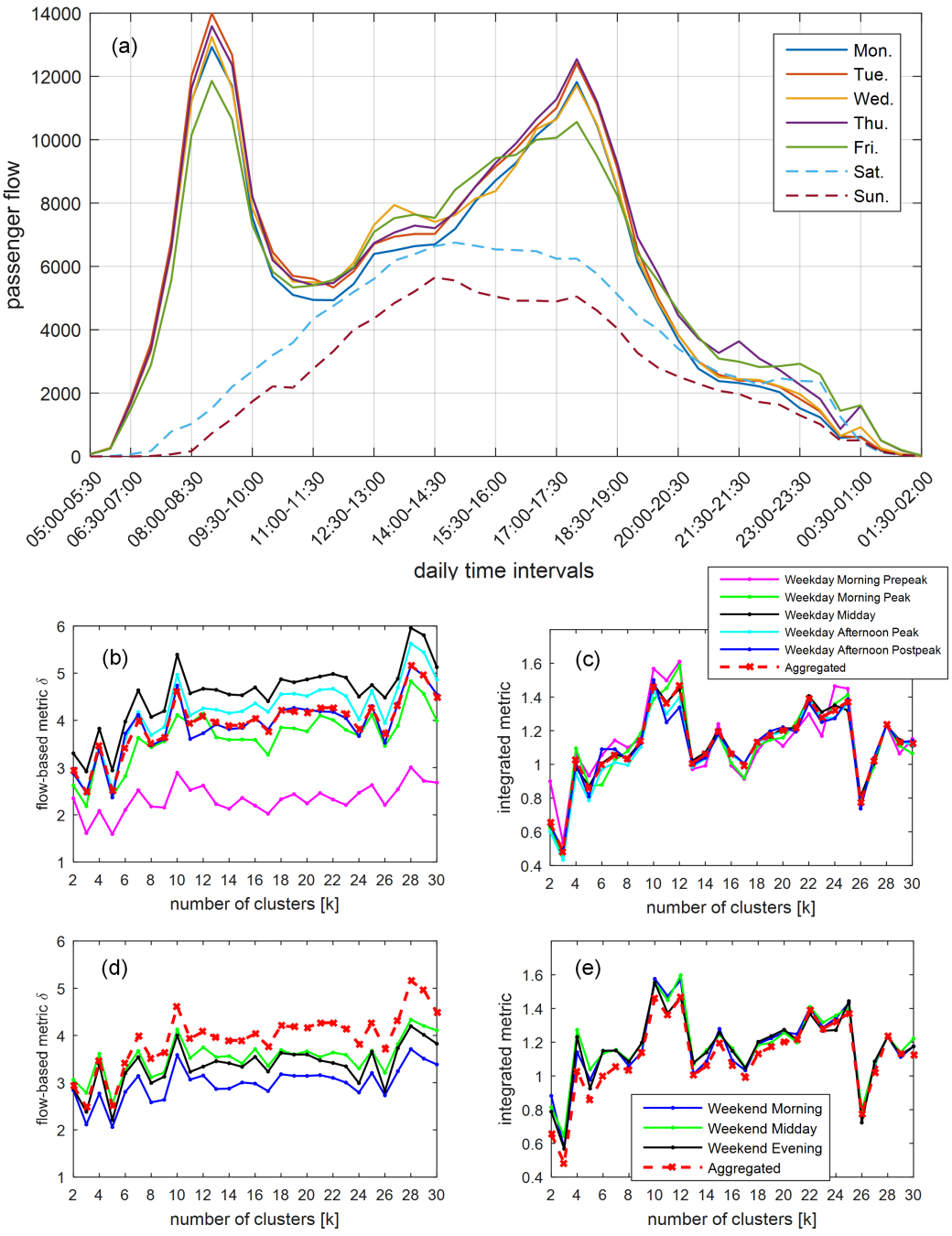


Figure 5.5: Temporal variability analysis. (a) within-day and across-day temporal PT demand; (b) time-dependent flow-based metrics for different periods over weekdays; (c) integrated metrics for different periods over weekdays; (d) time-dependent flow-based metrics for different periods over weekend; (e) integrated metrics for different periods over weekend;

Weekend

- Morning, before 10:00 a.m.;
- Midday, 10:00 a.m. to 6:00 p.m.;

- Evening, after 6:00 p.m.

Results for weekdays are shown in Figure 5.5b and Figure 5.5c, while those for weekends are in Figure 5.5d and Figure 5.5e. The red dash line plotted in all these figures represents the result of aggregated passenger flow over the entire study period and can be used as a benchmark. Overall, the temporal flow variance is found not to have a significant influence on the final determination of number of clusters. The best choices still remain in the neighborhood of $K11$, though $K12$ in some cases turns out to be optimal while $K10$ in the others. The general pattern remains stable.

One particular finding is that, during weekdays, the flow-based metric of midday is remarkably higher than the rest, while the one of morning pre-peak always remains the lowest. This implies that more long-distance inter-cluster journeys are generated when people are going to their workplaces early in the morning. At midday, on the contrary, the intra-cluster flow is stronger than the inter-cluster one because these traveling activities are less related to commuting. However, the metrics of afternoon peak and afternoon post-peak suggest a more mixed composition of journey purposes, such as performing shopping, recreational or household-related activities in the city after work. During the weekend, however, the flow-based metrics of all three periods stay at a low level, which implies that more inter-cluster journeys are performed than intra-cluster ones compared with weekdays. This can be explained by the fact that people normally go to the city for shopping or other leisure activities during the weekend.

5.5 Conclusion

Accurate estimation of passenger demand is crucial for both PT planning and operating processes. This chapter proposes a k-means-based station aggregation method which can quantitatively determine the clustering by considering both flow and spatial distance information. Differing from the traditional way of grouping stops based on TAZs, the proposed data-driven method provides another effective and efficient solution to those applications involving PT demand aggregation based on directly observed flows rather than their proxies. The method was specified and applied to a case study of The Hague, using the criterion that the ratio of average intra-cluster flow to average inter-cluster flow should be maximized while maintaining the spatial compactness of all clusters. This type of aggregation is particularly suitable for urban areas characterized by high density of PT stations, such as the case study area, The Hague. Travelers in such contexts are capable of choosing different origin and destination stations and services. The proposed method consists of four steps. Firstly, the best clustering of each K is constructed by running the k-means algorithms a number of times with different initial centers and selecting the one that results with the minimum SSE, a measure for the variance of clusters. Then two metrics based on distance and passenger flow are computed considering both intra-cluster and inter-cluster components. Finally, the two metrics are combined to determine the optimal number of cluster following the

criterion adopted. This analysis process can be applied using other spatial and flow metrics of interest, depending on the application, case study characteristics and data availability. The temporal variability analysis shows that the variance in passenger flow over time does not have a significant influence on the final determination of number of clusters when using the proposed method, which implies that this method is robust and can be potentially adopted for both short-term and long-term PT related research.

Regarding future research directions, an intriguing extension pertains to how the PT service network could also be considered in clustering. This will pave the way for multi-scale research on PT systems. In addition, choices on stops and routes within clusters and across clusters could also be investigated.

Chapter 6

Accessibility Analysis of Public Transport Networks

The previous three chapters have been focused on the research themes of information generation and passenger flows. In this chapter, we switch our attention to service networks, an important part of the supply side of PT systems. We present a new methodology for analyzing the accessibility of PT service networks using a network science approach. It allows for fast comparative assessment across multiple PT networks. Measuring the accessibility by average travel impedance, we propose an innovative weighted graph representation of public transport networks (PTNs) that explicitly incorporates travel costs according to the planned services contained in GTFS data. Consequently, the method enables efficient computation of minimal generalized travel costs between stop pairs. Such cost is comprised of initial and transfer waiting times, in-vehicle travel times and time-equivalent transfer penalty costs. To demonstrate the high transferability and efficiency of the proposed method, we present a case study assessing worldwide tram networks' accessibility. The results provide new insights into PTN design, benchmark and planning.

This chapter is organized as follows. An introduction is first presented in section 6.1 to describe the research motivation. Section 6.2 briefs related studies on PTN analysis and PT accessibility. The proposed methodology is then detailed in section 6.3. In section 6.4, the selected eight tram networks for the case study are described, followed by the presentation of the results and analysis in section 6.5. Section 6.6 concludes the study with the insights derived from the comparative study across different tram networks and suggestions for future research.

This chapter is an edited version of the following article:

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

6.1 Introduction

Network science, a research field built upon graph theory, is dedicated to studying the connection and interaction between components in complex systems (Newman, 2010). It provides researchers with powerful toolkit to quantitatively investigate the collective dynamics resulting from the interactions among system elements. Given this advantage, an increasing amount of research has been conducted to apply theories and methods from network science to study transport systems over the past decades (Lin & Ban, 2013). In particular, there has been a focus on studying the topological characteristics of PTNs (e.g., Sienkiewicz & Holyst, 2005; von Ferber et al., 2007, 2009; Berche et al., 2009; Louf et al., 2014; de Regt et al., 2019).

Applications of network science to transport systems have often been studied by network scientists seeking real-life examples of networks rather than by transport engineers and planners (Derrible & Kennedy, 2011). Consequently, a majority of studies ended up solely showing topological analyses without embedding information fundamental to transport systems, leading to claims that these works do not necessarily contribute to the transport community. For instance, Dupuy (2013) pointed out that these had provided limited recommendations to network planners, and thus impeded potential applications and impacts due to the absence of features related to transport and urban planning. Presumably, One of the main causes is that most topological analyses were performed with unweighted networks, thus leading to findings and conclusions that can provide limited knowledge and insights for improving the planning and operations of PT (Cats, 2017). Although the significance of analyzing weighted networks for more meaningful findings was already demonstrated almost two decades ago (Barrat et al., 2003), unweighted networks have still been commonly used in this specific research domain. Moreover, the perspective has often neglected the features associated with service attributes which are fundamental to PTNs. This is largely attributed to the difficulty in obtaining consistent PT attributes for creating meaningful weighted networks. This research gap needs to be bridged for network science to become more applicable in the transport research community.

To this end, this study is dedicated to performing an exemplary integration of network science and PT analysis that can enhance the understanding of system properties and performance. our primary contribution pertains to proposing a new method based on network science for computing public transport accessibility measured as the average travel impedance. Note that the concept “accessibility” in this study is defined solely based on the travel impedance associated with reaching any potential destination across the network. This definition thus does not account for the intensity and diversity of destinations, which stay beyond the scope of this study and are hence neglected in the remaining of this paper. The travel impedance metric is defined based on the generalized travel cost (GTC) between stop pairs, containing initial and transfer waiting times, in-vehicle travel times and time-equivalent transfer penalty costs. To perform efficient computation, a new type of weighted graph representation of public transport

networks is proposed, which explicitly incorporates the aforementioned components that are derived from GTFS data. The secondary contribution of this study consists in performing a comparative assessment of worldwide tram networks' accessibility. The analysis shows insights into how different travel components (e.g., in-vehicle travel times and waiting and transfer times) specifically contribute to the variance in accessibility across different networks. Such latitudinal comparative assessments can provide additional knowledge for the PTN design, benchmark and planning, but are still scarce in the current literature due to the requirements imposed by existing methods that heavily rely on geographical information systems (GIS).

6.2 Related research

The research on PTN analysis has for a long time been a significant topic among transport scholars given its fundamental role in influencing network design and planning (van Nes, 2002). According to a review by Derrible & Kennedy (2011), several chronological development stages of this research topic can be identified. Pioneering contributions were made early by Vuchic and Musso who introduced a graph theory approach for analyzing PTNs (Musso & Vuchic, 1988; Vuchic & Musso, 1991). A variety of metrics were established to quantitatively evaluate PTNs (Vuchic, 2005). Derrible & Kennedy (2009, 2010a,b) later on adopted graph theory principles for studying PTNs in a series of contributions. For example, they adapted a variety of concepts of graph theory to describe the features of metro networks. *State*, *form* and *structure* were proposed to measure the complexity of a network, the link between systems and the built environment, and the connectivity and directness of networks, respectively. The latest development has been focused on applying concepts, theories, and methods from network science to study the topological characteristics of PTNs. This has been largely facilitated more recently by the increasing availability of data sets (e.g., Gallotti & Barthelemy, 2015; Kujala et al., 2018a) and software functionalities (e.g., Hagberg et al., 2008).

6.2.1 Network science analysis of public transport networks

At the early stage, PTNs drew network scientists' attention when they searched for real-life examples of networks to test their theories and models (e.g., Latora & Marchiori, 2001, 2002; Sen et al., 2003). For example, many studies explored whether PTNs possess *scale-free* (Barabási & Albert, 1999) and *small-world* (Watts & Strogatz, 1998) features, which are considered among the most significant topological properties for characterizing networks. For more detailed summary and discussion of these studies, readers are referred to the reviews by Derrible & Kennedy (2011) and Zanin et al. (2018).

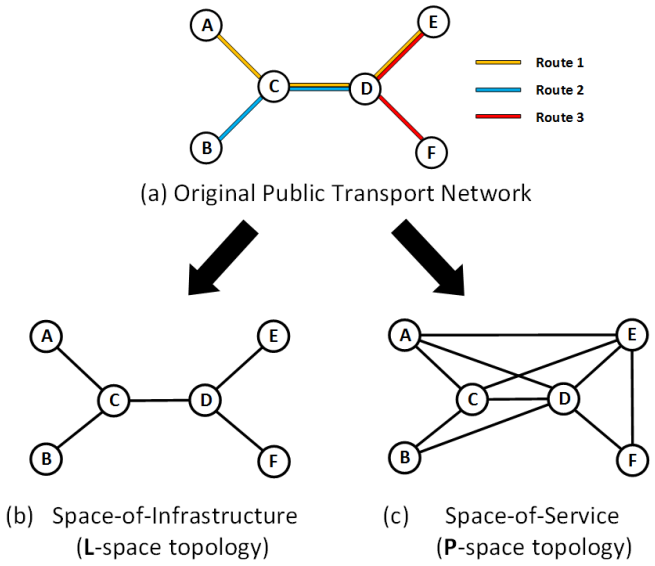


Figure 6.1: Illustration of two commonly adopted topological (space) representations of PTNs (adapted from von Ferber et al. (2009)). The terms *space-of-infrastructure* (**L**-space) and *space-of-service* (**P**-space) are used in the following to better reflect the context of PT.

Unraveling PTNs' topological characteristics requires considering the key feature of PTNs. In particular, PTNs consist of both infrastructure and service dimensions with the latter being superimposed on the former. Consequently, establishing meaningful graph representations of PTNs lays the foundation for any further network analyses. Two types of graph representations that have been mostly employed in the literature are the so-called **L**-space and **P**-space topology, which are illustrated in Figure 6.1 and also interpreted below.

- **L**-space: A straightforward representation of PTNs from the perspective of infrastructure. Each node represents a stop, and a link between two stops is formed if two stops are adjacent on at least one infrastructure segment (i.e. road or rail). This is one of the most extensively utilized representations by researchers (e.g., Latora & Marchiori, 2002; Sienkiewicz & Holyst, 2005; von Ferber et al., 2007, 2009). In many of the studies adopting this representation, only unweighted graphs are used in the analysis, hence containing no information of service properties.

- **P**-space: A representation solely based on PT service routes. The nodes represent stops and are linked if they are served by at least one common route. As a result, the neighbors of a node in this space are all stops that can be reached without performing a transfer. **P**-space has been widely used to investigate transfer possibilities (e.g., Sienkiewicz & Holyst, 2005; Xu et al., 2007; von Ferber et al., 2007, 2009). This space does not contain any information about the infrastructure layer (i.e., the physical sequence of links traversed between stops).

In the following, the terms *space-of-infrastructure* (**L**-space) and *space-of-service* (**P**-space) are used to better reflect the context of PT (Luo et al., 2019).

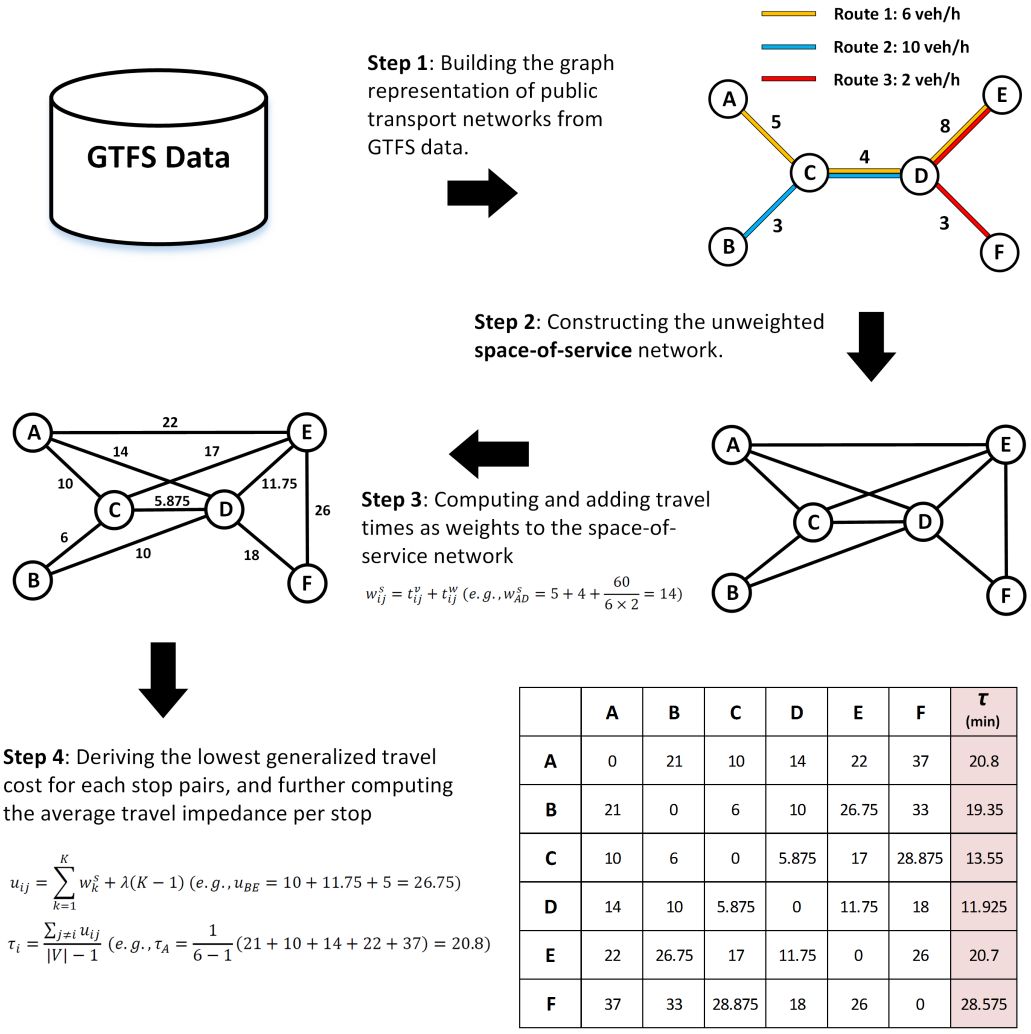
Based on the pioneering studies mentioned above, scholars have recently striven to better capitalize on network science in the context of PT research. This has led to efforts to incorporate PT specific features into complex network analysis of PTNs, such as travel demand and service attributes (e.g., passenger flows, transfers, service frequency, travel times, etc.). More weighted complex network analyses have emerged to account for demand and supply patterns in PTNs (e.g., Soh et al., 2010; Haznagy et al., 2015; Feng et al., 2017). Furthermore, investigations into the vulnerability, robustness and (node and link) criticality of PTNs have explicitly considered passenger demand and flow assignment (e.g., Cats & Jenelius, 2014; Cats, 2016; Cats et al., 2016, 2017).

6.2.2 Public transport accessibility

Accessibility has been widely studied in the field of urban planning (e.g., Batty, 2009) and transport planning (e.g., Geurs & van Wee, 2004; van Wee, 2016). According to the summary by Nassir et al. (2016), a consensus has been reached that accessibility can be measured based on two main components, which are: (i) locations and attractiveness of urban opportunities (benefit side); and, (ii) impedance of traveling to these locations from residential areas in the network (cost side). In a nutshell, more accessible areas are defined as those that can be reached with lower travel impedance. PT accessibility can thus be defined in a similar fashion, with the travel mode restricted to PT, and the travel impedance calculated based on PTN attributes.

PT accessibility research has so far been mostly performed based on GIS techniques with limited data availability (e.g., Lei & Church, 2010; Currie, 2010; Saghapour et al., 2016). The recent wide spread of GTFS data, along with the improvement of software's capability in processing them (e.g., ArcGIS), has further facilitated these GIS-based approaches (e.g., Farber et al., 2014, 2016; Farber & Fu, 2017). Moreoever, new analytical methods have also been gradually proposed in parallel, such as the travel impedance metrics based on utility (Nassir et al., 2016) and Pareto-optimal journey (Kujala et al., 2018b), and a faster computational algorithm by Fayyaz S. et al. (2017).

According to the literature review presented above, it can be summarized that previous studies have mostly performed mechanical analyses without incorporating service information fundamental to PT systems, thus providing the transport community with limited insights. Although some studies have attempted to address this challenge, more efforts are still needed. To further bridge the gap, the research presented in what follows demonstrates how network science can be effectively integrated with PT accessibility analysis.



	A	B	C	D	E	F	τ (min)
A	0	21	10	14	22	37	20.8
B	21	0	6	10	26.75	33	19.35
C	10	6	0	5.875	17	28.875	13.55
D	14	10	5.875	0	11.75	18	11.925
E	22	26.75	17	11.75	0	26	20.7
F	37	33	28.875	18	26	0	28.575

Figure 6.2: Illustration of the proposed method.

6.3 Methodology

This section presents the proposed method. An overview is first presented in Figure 6.2 along with an introduction of all the individual steps. The details of each step is then depicted in the following subsections.

6.3.1 Building graph representations of public transport networks from GTFS data

We first define that PTNs are comprised of two layers: the infrastructure layer (i.e., roads and rails) and the service layer superimposed on the physical one (i.e., routes). A PTN is then represented as a directed graph which can be denoted by a triple $G = (V, E, R)$, where V, E, R represent the set of nodes, links and routes, respectively. Each node v represents a stop, while each link e_{ij} is defined by an ordered pair of nodes (v_i, v_j), where v_i and v_j, respectively, denote the source and target nodes. Each route r

is directional and characterized by an ordered sequence of links $r = (e_{r,1}, e_{r,2}, ..., e_{r,|r|})$. Note that a link can constitute parts of different routes when they share the same corridors. In addition, the stop here relates to a service location (as commonly shown in PT maps) which can contain more than one individual boarding and alighting spot in the operational network.

Based on the above definition, PTNs are constructed from GTFS data which are now commonly available as a standard data format for PT operators to share their schedules and network information with the public. Structured as a relational database, one GTFS feed is comprised of multiple files that are interconnected via common keys (Google, 2019). To obtain required graph representation, a dedicated program was developed in MATLAB. In principle, a full-scan of all the trips associated with a route during a period needs to be performed in order to obtain the complete stop sequence for this route.

6.3.2 Constructing the unweighted space-of-service network

Based on the basic graph representation, we further establish the unweighted space-of-service network $G^s = (V^s, E^s)$ for PTNs. As introduced in section 6.2.1, this topological representation characterizes PTNs merely from a service perspective. A node v_i^s in this case represent a PT stop which is the same as it in the basic graph G (i.e., $V^s = V$). A link $e_{ij}^s \in E^s$ is created if v_j can be reached from v_i without performing transfers.

6.3.3 Adding travel times as weights to the space-of-service network

Using the scheduling information from GTFS data, we are able to combine in-vehicle travel times and waiting times into link weights in the space-of-service network. Note that both components are time-dependent based on schedules and are averaged values for a given time period. For simplicity the temporal element is not incorporated in the following formulation. For link e_{ij}^s in G^s, its weight w_{ij}^s is defined as follows:

$$w_{ij}^s = t_{ij}^v + t_{ij}^w \tag{6.1}$$

where t_{ij}^v denotes the in-vehicle travel time of the direct service from stop i to stop j according to the schedule from GTFS. t_{ij}^w represents the expected waiting time of travelers for the direct service from stop i to stop j during the same time period. This is estimated as half of the headway based on the joint service frequency of all direct services from stop i to stop j. It needs to be pointed out that if there is any synchronization among different PT routes, it is neglected in this definition. An example is presented at step 3 in Figure 6.2 to illustrate the proposed weight.

6.3.4 Measuring the average travel impedance

The average travel impedance associated with individual stop τ_i in this study is defined as follows:

$$\tau_i = \frac{\sum_{j \neq i} u_{ij}}{|V| - 1} \tag{6.2}$$

where $|V|$ denotes the number of stops in the PTN. u_{ij} represents the lowest GTC from stop i to j. Specifically, u_{ij} is the sum of total in-vehicle travel times, total waiting times for each journey leg (i.e., the initial waiting time and subsequent transfer waiting times), and time-equivalent penalty per transfer. The last component expresses the additional penalty associated with the inconvenience of performing a transfer beyond the additional travel time. The lowest GTC in this case can be efficiently computed using the proposed weighted space-of-service network $G^s(V^s, E^s, W^s)$. For instance, for the GTC from stop i to j, u_{ij}, the shortest path P_{ij} between them in G^s is first searched for, which is denoted as a sequence of links $P_{ij} = (e^s_1, e^s_2, ..., e^s_K)$. Then the additional penalty cost is added based on the number of transfers along this shortest path. The definition is specified as follows:

$$u_{ij} = \sum_{k=1}^{K} w^s_k + \lambda (K - 1) \tag{6.3}$$

where w^s_k denotes the weight (i.e., travel time , including in-vehicle time and waiting time) of link e^s_k. K represents the number of links, namely the minimal number of journey legs from stop i to j. λ represents the time-equivalent penalty cost per transfer which is a constant in this case. The applied metric is conceptually similar to the so-called *closeness* centrality (Bavelas, 1950), but it is much more adapted in the context of PT. It allows to quantify the travel impedance of each stop to all other stops in terms of the total journey time in the PTN. The simple average function can be replaced by a weighted average to enrich the metric by accounting for the importance of different origin-destination relations based on land-use or demand data if those are available.

6.4 Case study: assessing the accessibility of tram networks

To demonstrate the proposed method, eight tram networks worldwide, including Melbourne (Australia), Vienna (Austria), Milan (Italy), Toronto (Canada), Budapest (Hungary), Zurich (Switzerland), Amsterdam and The Hague (The Netherlands), are employed for the case study. Tram networks are used because there is, to the best of our knowledge, a lack of dedicated research on tram networks with sufficient samples in the literature compared to other PT modes, particularly metros (e.g., Derrible, 2012;

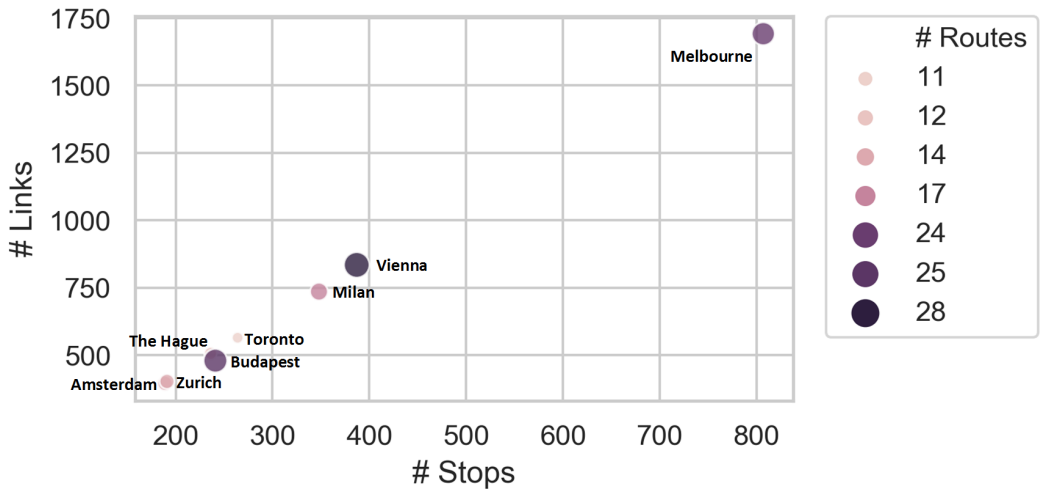


Figure 6.3: Illustration of the basic properties of the studied tram networks. Note that the stop here relates to a service location (as commonly shown in PT maps) which can contain more than one individual boarding and alighting spot in the operational network

Roth et al., 2012). This was presumably largely due to the unavailability of standardized data on tram networks. Therefore, by investigating eight tram networks worldwide in this case study, we aim to further highlight one of the main merits of the proposed method, which is its high transferability.

The network selection criteria include the following aspects: (i) the tram network plays a dominant role in the local urban transport system; and (ii) the GTFS data are available. All the networks were then generated using the GTFS data from the second half of 2018, which are subject to temporal variations possibly caused by construction and maintenance. In addition, for this case study the weekday morning peak schedules (8 am – 9 am) are used for computing the weight of space-of-service networks. For the time-equivalent transfer penalty cost λ in Equation 6.3, we apply 5 minutes per transfer in this study based on the result from Yap et al. (2018).

To obtain a better understanding of the basic properties of these tram networks, a graph displaying the numbers of stops, (physical) links and routes (i.e., $G(V,E,R)$) is presented in Figure 6.3. Noticeably, the stop here relates to a service location (as commonly shown in PT maps) which can contain more than one individual boarding and alighting spot in the operational network. From the graph, it can be seen that these eight tram networks exhibit considerable differences in terms of their physical scale. With more than 800 stops and 1500 links, the Melbourne tram network is significantly larger than the others. Notwithstanding, it does not contain the largest number of routes (24 routes only). Following the world's largest tram network in terms of the track-km length, the second tier cluster in our analysis encompasses Vienna and Milan, both of which possess some 300 stops and 700 links. However, Vienna boasts the largest number of regular routes out of all the eight networks, while Milan's number of

routes seems to be proportionate to its scale. On the lower left corner in this diagram lie the remaining five networks, with Amsterdam and Zurich being the smallest. The Budapest tram network also stands out since its number of routes is remarkably disproportionate to its scale. Overall, it can be concluded that the numbers of stops and links are linearly correlated, while the number of routes does not necessarily coincide with the size of the networks.

6.5 Results and analysis

This section presents the results of applying the proposed method to the case study tram networks. Section 6.5.1 first introduces a benchmark metric used in the analysis. Then the results are presented in section 6.5.2 with visualizations, followed by a discussion in section 6.5.3. More insights are further presented in section 6.5.4 through an investigation into the variance in the proposed GTC-based travel impedance metric across different networks.

6.5.1 Additional benchmark travel impedance metric

To better demonstrate the merit of the proposed GTC-based travel impedance metric, especially in the context of PT, we include an additional one in this analysis as a benchmark. It is computed using the space-of-infrastructure (**L**-space) representation described in section 6.2.1. This metric is derived in the same way as implied by equation 6.2, except that the GTC is replaced with the topological shortest path length in the unweighted space-of-infrastructure network. In other words, the travel impedance is solely determined based on the minimal number of links that are traversed between stop pairs in the space-of-infrastructure network. Neither travel times nor transfer attributes are incorporated in this benchmark metric. The travel impedance is thus completely viewed from a topological perspective. The reason for adopting this specific metric as the benchmark against the proposed one is the following: existing topological analyses of PTNs have used the space-of-infrastructure representation to study PTNs' properties related to what this study is examining. For example, several studies have studied the efficiency of PTNs based on the shortest path length in the unweighted space-of-infrastructure network without taking into account of PT transfers (e.g., Latora & Marchiori, 2001; de Regt et al., 2019). By comparing the proposed metric based on the GTC to this benchmark, the impact of including information on service properties can be assessed.

6.5.2 Results

In this subsection, the computational results are presented in the form of travel impedance maps with explanations. As Figure 6.4 shows, each column displays three graphs for a

given tram network, corresponding to the result of, from top to bottom, (i) the benchmark metric; (ii) the proposed GTC-based metric, and; (iii) the difference between (i) and (ii). For the first two graphs, the same color scheme is used to illustrate different magnitudes of the travel impedance, therefore allowing for analyzing the spatial variation in impedance. Additionally, the distribution of metrics is presented in the top right corner for a more quantitative description. Note that there are black "**x**" markers in the middle and bottom rows, which represent the stops disconnected from the rest of the network. This is due to the fact that in this study the weighted space-of-service network pertains to service provision rather than infrastructure availability and is thus time-dependent. Besides, there is a tram route that is completely disconnected from the rest of the network in Toronto. This is because this tram route is integrated into the network through two other metro routes which are not included in this analysis.

The graphs on the bottom row in Figure 6.4 visualizes the gap between the two different travel impedance metrics. The gap is quantified by the residual of a linear regression model in which the dependent and independent variables are the GTC-based and benchmark metrics, respectively. The rationale is that the bigger the residual (absolute value) for a stop is, the larger the gap between the two metrics for this stop is. For the network-wide visualization, we apply two colors, i.e., red and blue, to indicate the sign of residuals, while the magnitude is reflected through the depth of the color: the deeper the color is, the larger the gap between the two metrics is. Specifically, the blue spots illustrate where the travel impedance by the benchmark metric is higher than that by the proposed GTC-based metric, while the opposite holds for the red spots. The scatter plot in the upper right corner further shows more information about the linear regression between the two metrics. With the color for describing the gap retained, this plot allows for a more intuitive understanding of the consistency between the two metrics, particularly across different networks with the indication of the Pearson correlation coefficient r.

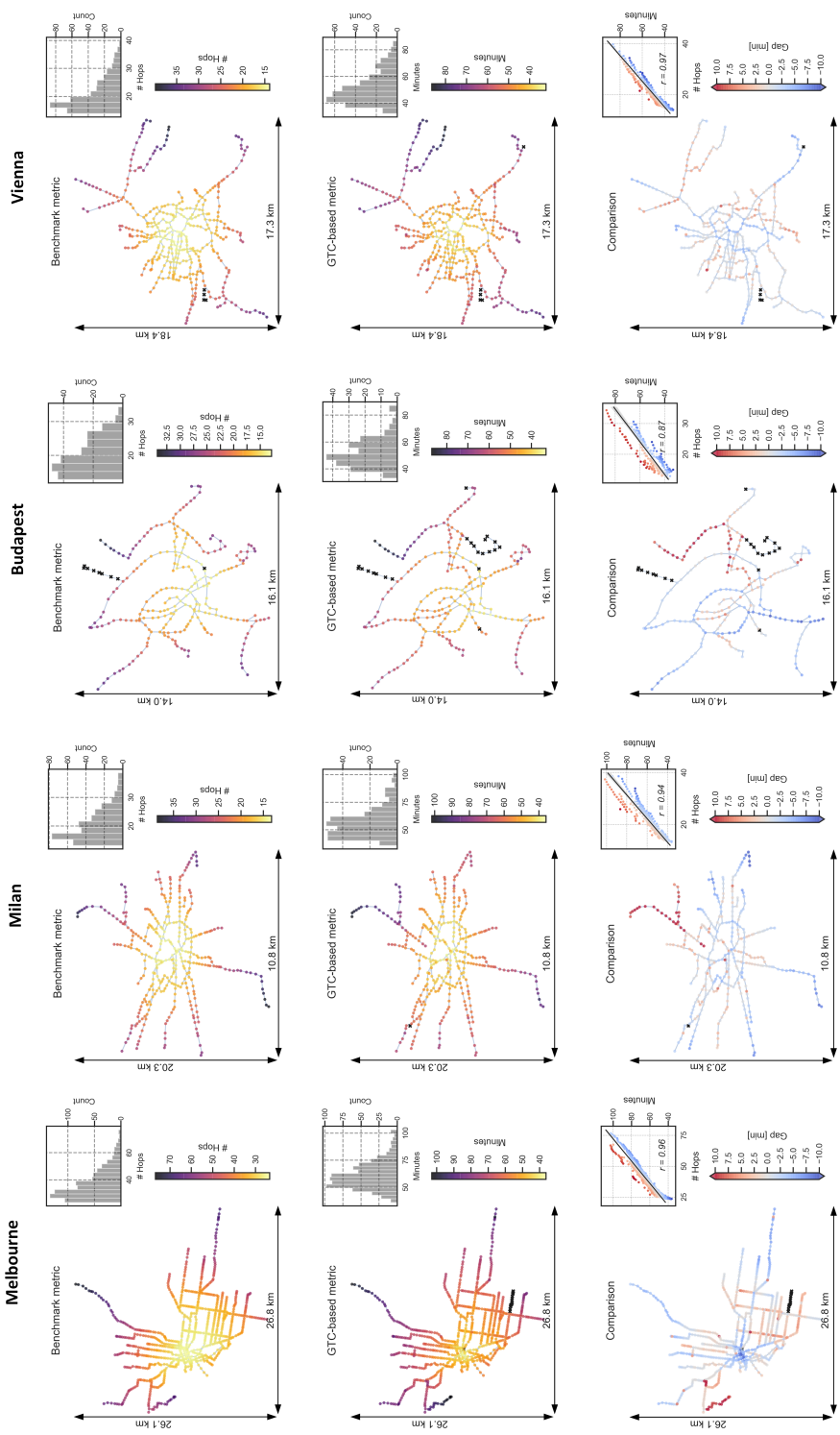


Figure 6.4: Visualization of the travel impedance maps for case study tram networks. The benchmark metric, newly proposed GTC-based metric, and the comparison between them are respectively displayed from top to bottom for each city. The physical scale of all the networks are also provided on axes.

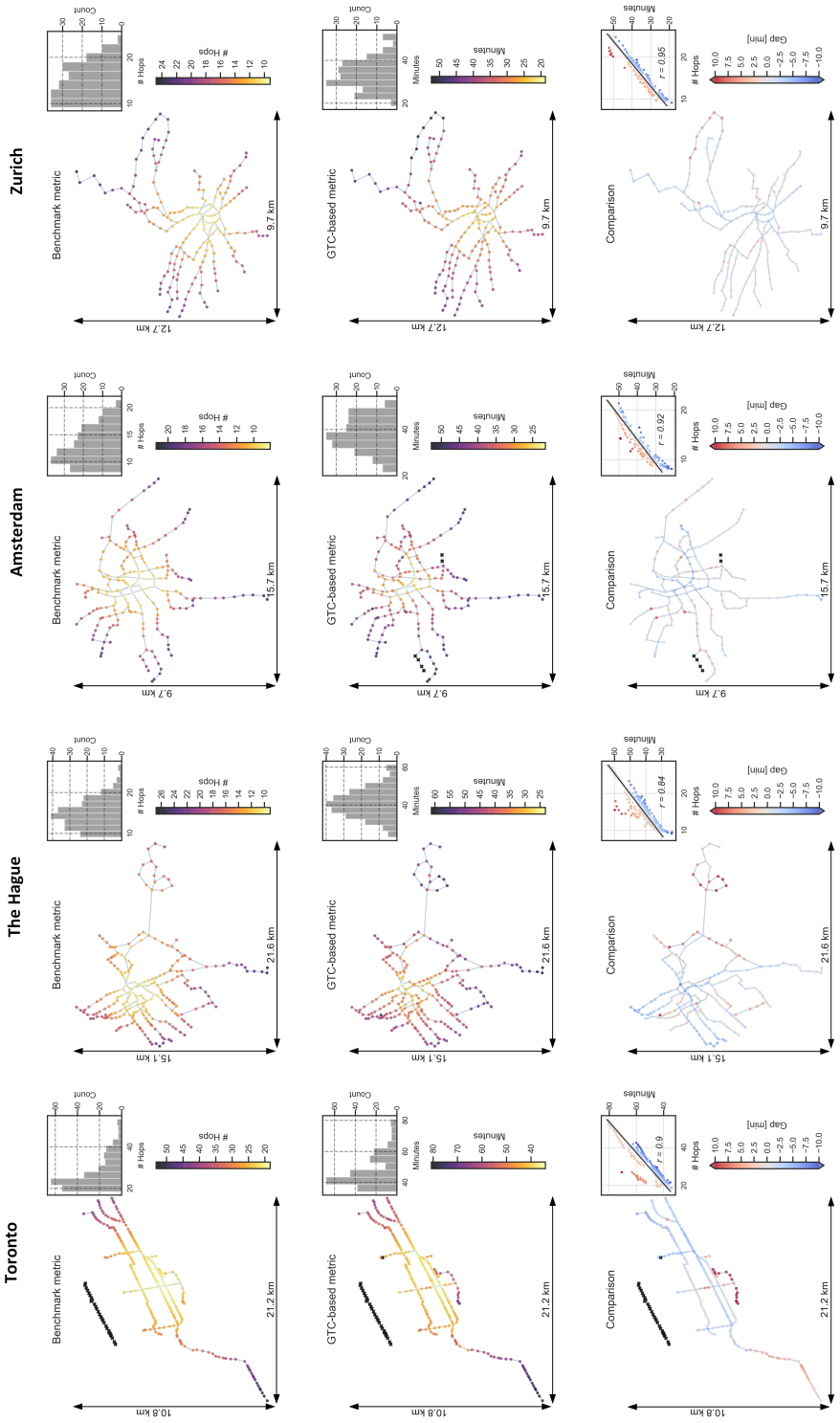


Figure 6.4: Continued.

6.5.3 Discussion

Some important patterns from the visualization are summarized in this subsection. First, according to Figure 6.4, it can be seen that low travel impedance (i.e., high accessibility) is pronounced for those stops in the central area of the network in both benchmark and GTC-based cases. The travel impedance gradually increases as one moves away from the center toward the periphery for all the networks. Nevertheless, certain differences are also noticeable pertaining to the general pattern and the distribution. For example, the spatial disparity in the travel impedance is visually more remarkable when it is measured in terms of the GTC (the second row). In other words, in the GTC-based case, the stops with low travel impedance are more concentrated in the central part of the network. This is particularly remarkable in the case of Amsterdam, The Hague, Vienna and Melbourne. Second, in several cities, e.g., The Hague, Toronto, Milan, Budapest and Melbourne, some stops on the periphery of the network display largely underestimated impedance by the benchmark metric (reflected by the red spots). These patterns pertaining to the inconsistency between the two metrics can stem from three factors: (i) Less waiting times are needed in the central area given higher service frequency, particularly in the studied morning peak period; (ii) It is more likely that reaching a stop distant from the center requires more transfers, thus incurring more transfer penalty costs. (iii) It is common that the physical inter-stop space increases as the location moves from the center to the periphery of a network, therefore requiring more in-vehicle travel times.

Some intriguing findings can be also observed from the Pearson correlation coefficient. It can be seen that the values of Vienna (0.97), Melbourne (0.96) and Zurich (0.95) are higher than those of the rest of the tram networks, implying a good consistency between the two metrics. The Hague (0.84) turns out to be significantly lower than any other network, followed by Budapest (0.87). In the case of the former, this is due to a cluster of suburban stops on the east since the long corridor connecting the suburb to the city largely increases the cost in the GTC-based situation, resulting in a bias in the benchmark metric which underestimates the impedance.

We further analyze the difference between the two metrics from a distributional perspective. From Figure 6.4, it can be already observed that the distributions of the two impedance metrics are significantly different from each other. The distribution of the benchmark metric is positively skewed, meaning that a large proportion of the distribution mass lies on the left side and there is a long right-side tail. While for the GTC-based metric, the majority in the distribution significantly moves to the right for most networks. Overall, the GTC-based metric quantifies the travel impedance in a less extreme way than the benchmark one does, suggesting that the spatial disparity in accessibility might be lower than assumed when neglecting service properties.

A more global comparison between them is further presented in Figure 6.5 using the complementary cumulative distribution plot. Figure 6.5a shows that several cities clearly overlap when the impedance is measured by the benchmark method (e.g.,

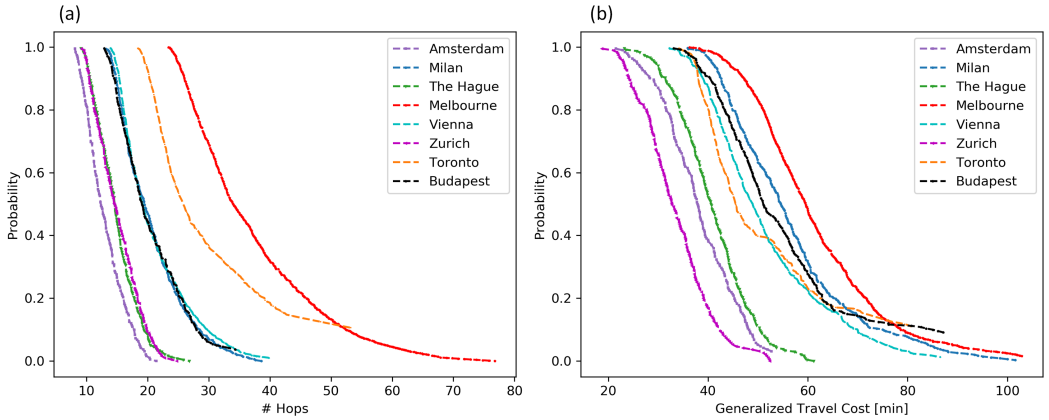


Figure 6.5: Complementary cumulative distributions of the travel impedance for the studied tram networks. (a) The benchmark metric; (b) The GTC-based metric.

the group of Budapest, Vienna and Milan; the group of Amsterdam, The Hague and Zurich), while Figure 6.5b demonstrates that the difference in the impedance among these networks becomes pronounced when the GTC is considered. In other words, once accounting for service properties that reflect network design and resource allocation choices, there are more pronounced differences between different networks than if those are neglected. Another finding is that the relative positions of the curves representing Budapest, Vienna and Milan shift from Toronto's left to its right when the GTC-based metric is applied. This could be attributed to the fact that the number of routes in Toronto is significantly lower than the other three, albeit the similar network size in terms of the number of nodes. As a result, much fewer transfers need to take place in Toronto than in the others, hence incurring lower transfer penalty costs.

6.5.4 Variance analysis

This subsection is devoted to investigating the variance in the resulting GTC-based travel impedance metrics across different networks. In addition to looking at the variance in the total cost, we also examine how individual components (i.e., (i) in-vehicle travel times and (ii) waiting and transfer times (with penalty)) varies in different networks. The so-called violin plot is used for visualization as shown in Figure 6.6. As a type of enhanced visualization technique for the conventional box plot (The median of the data is marked by a white dot inside the plot with the interquartile ranges lying on its both sides), the violin plot for each network additionally illustrates the probability density of the data smoothed by a kernel density estimator. In this case, the individual violin plots are arranged in a descending order from left to right in terms of the median travel impedance value of each city.

Figure 6.6a well corresponds to the intuition that larger networks in terms of the spatial scale indeed exhibit higher travel impedance. It can be observed that the median value of the top one (Melbourne, 60 min) is almost twice as that of the bottom (Zurich, 30

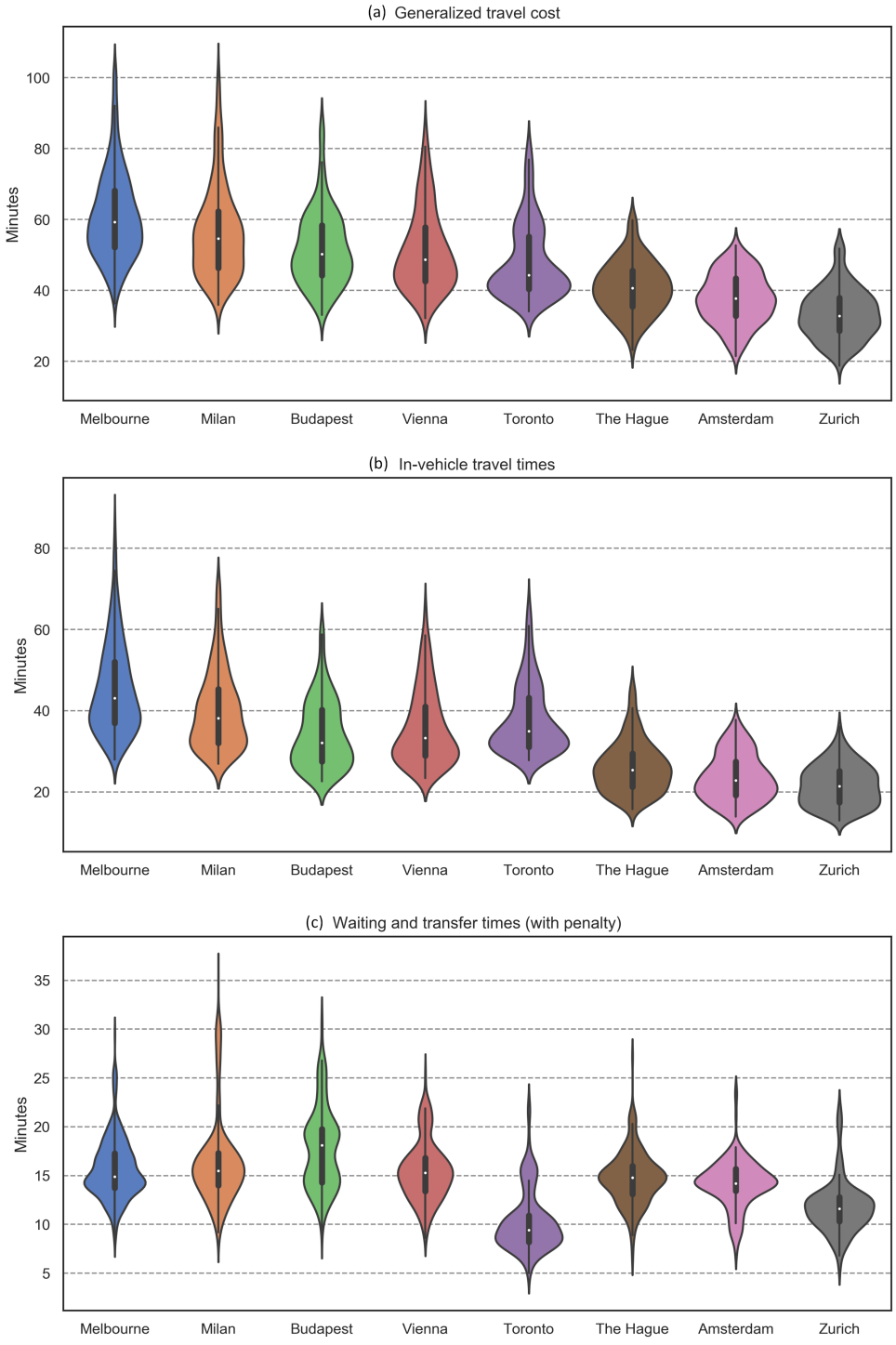


Figure 6.6: Visualization of the variance in travel impedance for eight tram networks. The violin plot displays the median, quartiles and probability density of the data smoothed by a kernel density estimator.

min). In addition, the diagram reveals that the variance in travel impedance is proportional to the spatial scale as well. As the median value of the impedance metric declines, the variance also drops in a way that the range between the maximum and minimum shrinks. For large networks, such as Melbourne and Milan, they appear

to have long tails on the top (hence, right-hand tails) as a result of dramatic network sprawl from the center to suburbs. The opposite holds for some much smaller networks, including The Hague, Amsterdam and Zurich. Their shapes look much more compact in comparison to the others. Besides, the unique shape of Toronto reveals that its tram network is differently organized from many others.

Figure 6.6b and Figure 6.6c further unravel how (i) in-vehicle travel times and (ii) waiting and transfer times (with penalty) respectively contribute to the variance in the GTC. It can be seen in Figure 6.6b that the sequence derived based on the total GTC can still be roughly applied to the case of in-vehicle travel times and the shapes are also similar. This makes sense as the overall GTC is dominated by the in-vehicle travel time in all the networks. The only remarkable inconsistency occurs to Toronto where its median value is larger than that of Budapest and Vienna.

Figure 6.6c then displays a very different pattern from the previous two diagrams. By excluding the component of in-vehicle travel times, which is largely dependent on the spatial scale of networks, we can actually find out how planned services on top of the physical network (i.e., route and frequency design) influence the travel impedance. This can be explained from two aspects: (i) The number of transfers needed for connecting every stop pairs in the entire network determines waiting times and transfer penalty costs; (ii) Service frequency determines waiting times for each ride. It can be seen that the majority of the networks exhibit a median value of around 15 minutes, with the largest and smallest being close to 20 minutes (Budapest) and lower than 10 minutes (Toronto), respectively. Surprisingly, most significant variance can be found in Milan which shows a long tail on the top. This explains why Melbourne and Milan are similar in terms of the shape of the GTC (Figure 6.6a) despite the fact that the in-vehicle travel times in Milan have less variance.

6.6 Conclusion

Existing complex network analyses of PTNs have mostly been focused on performing topological analyses without incorporating information fundamental to public transport services for the sake of PT planning. Although these studies have generated many insights into PT systems previously unknown to PT scholars, more efforts are still needed to enable network science to further address challenges encountered in PT planning and operations. To this end, this study proposes an exemplary integration of network science and public transport accessibility analysis. The primary contribution lies in developing a method based on network science for computing public transport accessibility measured as the average travel impedance. With the proposal of an innovative weighted graph representation of PTNs that explicitly incorporates travel costs according to the planned services contained in GTFS data, we are able to efficiently compute the minimal generalized travel cost between stop pairs. Such cost is comprised of initial and transfer waiting times, in-vehicle travel times and time-

equivalent transfer penalty costs. The secondary contribution pertains to performing a comparative assessment of worldwide tram networks' accessibility based on the proposed method. Such latitudinal comparative assessments can provide additional insights into the PTN design, benchmark and planning, but are still scarce in the current literature due to the requirements imposed by existing methods that are heavily reliant on GIS.

New insights derived from the comparative case study are summarized as follows. According to the comparison between the benchmark and proposed GTC-based metrics, the main conclusion is that the spatial disparity in PT accessibility can be higher when planned service properties (i.e., travel times, initial and transfer waiting times, and transfer penalties) are considered than when they are not. In addition, the subsequent investigation shows that larger networks in terms of the spatial scale indeed exhibit higher median total travel impedance. However, this is primarily attributed to the component of in-vehicle travel times which is intrinsically positively correlated to the spatial scale of networks. By excluding in-vehicle travel times, we can see that the majority of the networks exhibit a median cost of around 15 minutes regarding waiting and transfers, with the largest and smallest being close to 20 minutes (Budapest) and lower than 10 minutes (Toronto), respectively. Among all the studied networks, Milan shows the largest variance in this specific cost component.

We point out two directions for future research. First, OD demand can be incorporated when computing the average travel impedance. This would result in a more comprehensive way of measuring PT accessibility. Second, the proposed method can be extended to be more suitable for multimodal PT systems by adopting the multilayer network theory (Kivelä et al., 2014).

Chapter 7

Passenger Flow Modeling based on Network Properties

Having studied passenger flows and service networks separately in the previous chapters, we are further motivated to examine the relation between them from a data-driven perspective. To this end, this chapter tries to answer whether passenger flow distribution can be estimated solely based on network properties in PT systems. We use concepts and methods from network science, including the topological representation of PT infrastructure and service networks and centrality indicators, to systematically and concisely quantify PT network properties. Moreover, we offer interpretations of the applied centrality indicators in the context of PT systems. We then use panel data models given the dependencies across time and space. The results show that the selected network properties can indeed be used to approximate the global passenger flow distribution across the network to a reasonable extent of accuracy using solely regression models, although no causality should be implied. This finding can lead to the development of (i) parsimonious data-driven PT assignment models and (ii) global metrics for overseeing large-scale flow dynamics and the gap between flows and service supplies.

This chapter is structured as follows. Section 7.1 describes the context and motivation of the research, followed by the explanation of the proposed methodology in 7.2. Section 7.3 describes the case study networks and experimental setup. The results and discussion are presented in section 7.4. The chapter ends with the conclusions drawn in section 7.5.

This chapter is an edited version of the following article:

Luo, D., Cats, O. & van Lint, H. (2019) Can passenger flow distribution be estimated solely based on network properties in public transport systems? *Transportation*. https://doi.org/10.1007/s11116-019-09990-w.

7.1 Introduction

Estimation and prediction of passenger flow distribution is one of the most significant topics in the field of PT research given its critical role in assisting planning and management. The conventional approach, like that in the road traffic research, is to develop passenger assignment models which take demand profiles – typically in the form of origin-destination matrices – as input and then distribute the demand across the network (Ortúzar & Willumsen, 2011). These models are normally referred to as *transit assignment models* in the transport research community, and their core pertains to modeling travelers' route choices in PT systems as functions of network conditions and travel preferences (Liu et al., 2010). Two types of static equilibrium transit assignment models have been mostly developed over the past decades, namely the frequency-based and schedule-based. The major distinction between them lies in the representation of PTNs given their substantial impact on the passenger loading procedure (Gentile et al., 2016). More specifically, the frequency-based approach represents PTNs at the route-level with corresponding frequencies (e.g., Nguyen & Pallottino, 1988; Spiess & Florian, 1989; Cepeda et al., 2006; Schmöcker et al., 2011), while the schedule-based one enables a more detailed representation of time-dependent specific vehicle runs (e.g., Nuzzolo et al., 2001; Zhang et al., 2010).

Notwithstanding the continuous development of transit assignment models, other possibilities for understanding and further modeling the passenger flow distribution in PT systems to a network-wide extent have remained underexplored. Presumably, this is a result of the longstanding data scarcity in the field. Under this "data-poor and assumption-rich" situation (Vlahogianni et al., 2015), the conventional modeling approach has undoubtedly provided the most feasible solution to this challenging problem. Nonetheless, as the capability in measuring the PT passenger flow dynamics in a large spatiotemporal scale becomes increasingly available owing to emerging PT demand data sources, such as the automatic fare collection data (Pelletier et al., 2011), it is now worth investigating whether there can be alternative ways to model the passenger flow distribution in PT systems.

This study hence examines a research question: Can passenger flow distribution be estimated solely based on network properties in PT systems? While the answer to this question looks apparent, it has not been empirically investigated to a sufficient extent. One can make an underlying assumption that transport network properties should of course correspond to passenger flow distribution since networks are supposedly designed to efficiently accommodate prevailing demand patterns in PT systems (van Nes et al., 1988). However, it shall be stressed that a range of other factors, such as travelers' behavior, historical network development and physical constraints, also have non-negligible influences on demand and network structure in any transport systems. In fact, the discussion about whether traffic flows can be approximated by network properties in urban street networks has lasted for decades among urban planning researchers (e.g., Hillier et al., 1993; Penn et al., 1998; Turner, 2007; Jiang & Liu, 2009;

Kazerani & Winter, 2009; Gao et al., 2013). Recent evidence was provided by Gao et al. (2013) based on the traffic volume derived from the GPS-enabled taxi trajectory data from a Chinese city. Their study concludes that the betweenness centrality, which has been commonly employed as a local indicator of network properties, is not a good predictive variable for urban traffic flow and the gap can be explained by the spatial heterogeneity of human activities and the distance-decay law. In addition, a limited amount of research attempts have also been made by scholars from various fields in the past few years to examine the relation between network properties – mostly limited to the betweenness centrality – and the traffic flows in urban road traffic systems. (e.g., Altshuler et al., 2011; Puzis et al., 2013; Ye et al., 2016; Zhao et al., 2017b; Wen et al., 2017; Akbarzadeh et al., 2019). No such comparable effort, however, has been made in the context of PT systems, which therefore necessitates dedicated investigations into the proposed research question above.

To this end, we conduct this study with the methodology developed in a reverse engineering fashion, which unravels the correlative relation between passenger flow distribution and network properties in PT systems. Differing from previous studies, we examine a variety of network properties by considering the centrality indicators in different topological representations of PTNs. We show how concepts originating from the domain of complex network science can be applied and interpreted in the context of PT systems. We further apply the proposed methodology to two real-world tram networks in The Netherlands, i.e., The Hague and Amsterdam, where passenger flow observations are available. Regression models capturing the correlation between passenger flow distribution and several centrality indicators are first developed using the data from The Hague, and are then evaluated for both networks separately. Note that no causality is implied by the models. It is the correlation - rather than the causation - between PTN properties and passenger flow distribution that is essentially investigated. Moreover, the unraveled relation and developed models have the potential to serve as a complementary tool for PT operations management, while it is inappropriate to apply them to the long-term passenger flow forecasting.

7.2 Methodology

7.2.1 Overview

An overview of the research structure is shown in Figure 7.1 with the workflow and components of the methodology sketched. In the beginning, the representation of PTNs is described in section 7.2.2 to lay the foundation. Then, section 7.2.3 and 7.2.4 are respectively dedicated to illustrating the independent and dependent variables in this study, namely centrality indicators of PTNs and passenger flow distribution, both of which are considered in a time-dependent manner. The model development is later described in section 7.2.5, followed by the model evaluation in section 7.2.6.

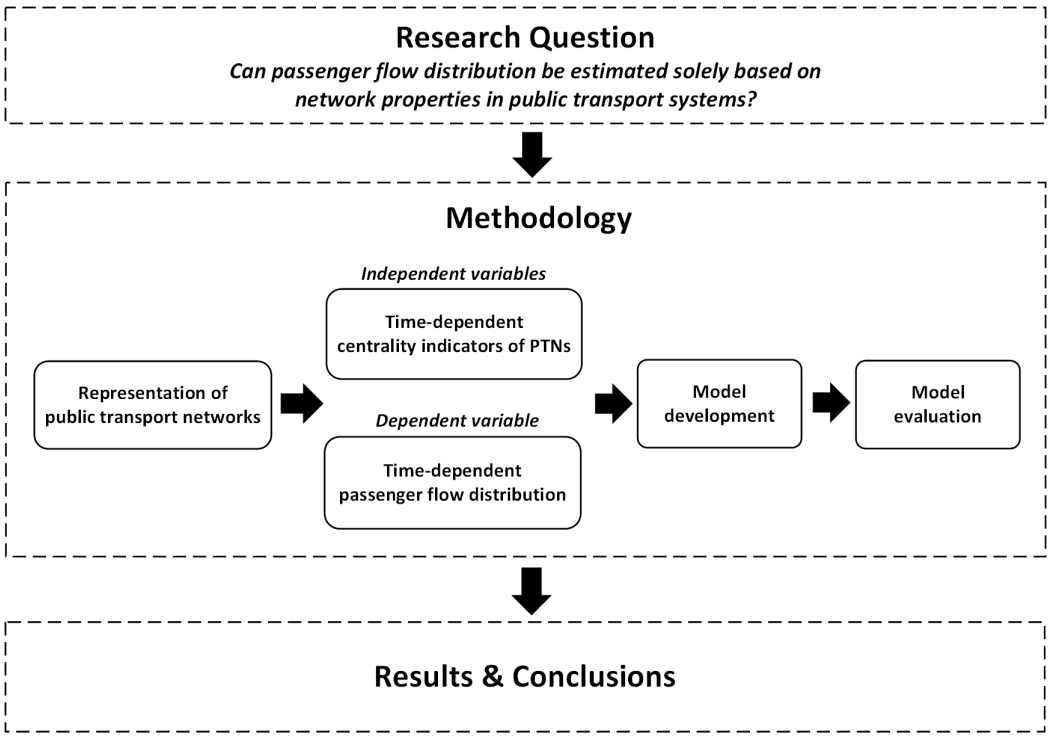


Figure 7.1: Illustration of the overall research design and the components and pipeline of the developed methodology.

7.2.2 Representation of public transport networks

We first clarify that the term "public transport network" in this study is referred to as the combination of two layers: the infrastructure network (i.e., road and rail) and the service network superimposed on the physical layer (i.e., routes). We then define a PTN as a *directed graph* with a triple $G = (V, E, R)$, where V, E, R represent the set of *nodes, links, routes*, respectively. Each node $v \in V$ represents a stop, while each link $e \in E$ is defined by an ordered pair of nodes (u, v), where u and v, respectively, denote the *source* and *target* nodes. Each route $r \in R$ is characterized by an ordered sequence of stops $r = (v_{r,1}, v_{r,2}, ..., v_{r,|r|})$ as well as an ordered sequence of links $r = (e_{r,1}, e_{r,2}, ..., e_{r,|r|})$. Note that a link can be utilized by multiple routes, and the direction of a route is also distinguished. In addition, the stop in this definition refers to a service location which can contain more than one individual boarding and alighting spot in the operational network.

Based on the fundamental representation of PTNs, we further apply two topological representations, the **L**- and **P**- space (von Ferber et al., 2009), to characterize the topology of PTNs' two different layers, i.e., infrastructure and service. These topological networks, which can be represented by adjacency matrices, are suitable inputs for further analyses. As Figure 7.2 illustrates, the **L**-space is a straightforward representation of PTNs' physical infrastructure. Each node represents a stop, and a link between two stops is formed if two stops are adjacent on an infrastructure segment (i.e. road or rail). Moreover, duplicate connections between nodes are not allowed. The **P**-space is

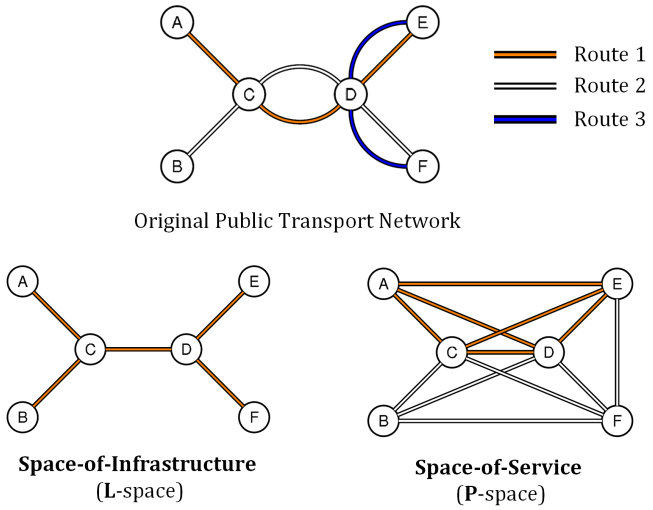


Figure 7.2: Illustration of the **L**-space and **P**-space representations of the exemplary PTN on the top, which consists of three routes and six stops (adapted from von Ferber et al. (2009)). The **L**-space essentially represents the infrastructure layout, while the **P**-space characterizes the PT service layer: stops that are directly linked require no transfer to reach each other. In order to make the use of these two topological representations more intuitive in the context of this study, we replace the term "**L**-space" and "**P**-space" with "*space-of-infrastructure*" and "*space-of-service*" in the remaining of this chapter.

constructed solely based on the service layer designed by PT operators/agencies, i.e., routes. The nodes in this space also represent stops, and two nodes are linked if they are served by at least one common route. In this sense the neighbors of a node in this space are all stops that can be reached without performing a transfer. In order to make the use of these two topological representations more informative in the context of this study, we replace the terms "**L**-space" and "**P**-space" with "*space-of-infrastructure*" and "*space-of-service*" in the remaining of this chapter.

Further enrichment of the topological networks of PTNs is performed by adding link weights related to PT service attributes. The space-of-infrastructure is enriched in two ways, including the in-vehicle travel time as a type of link cost and vehicle frequency per time unit as a type of link importance.With common routes considered, the weight of a link's ultimate frequency is determined by summing up the frequencies of all the routes traversing it, i.e., labeling the link with the respective joint frequency, which is consistent with the definition of space-of-infrastructure representation. For the space-of-service, the expected waiting time for a PT vehicle during a given time slice is considered as a type of link cost, which is defined as half of the planned headway with joint vehicle frequency between stop pairs considered. This definition is based on the assumption that (i) passenger arrival at stop is random in the context of urban high-frequency services, and (ii) arrival times of vehicles serving different lines is independent, i.e. no systematic synchronization is performed in the context of urban high-frequency services. Both unweighted and weighted topological networks will be

Table 7.1: Summary of the centrality indicators used in this study.

PTN representation	Notations	Centrality indicators	Weight	Weight attributes
Space-of-Infrastructure	$\mathbf{d}^{\mathbf{L},+/-}$	In/Out-degree	✗	–
	$\tilde{\mathbf{d}}^{\mathbf{L},+/-}$	In/Out-degree	✓	Vehicle frequency
	$\mathbf{b}^{\mathbf{L}}$	Betweenness	✗	–
	$\tilde{\mathbf{b}}^{\mathbf{L}}$	Betweenness	✓	In-vehicle travel time
	$\mathbf{c}^{\mathbf{L},+/-}$	In/Out-closeness	✗	–
	$\tilde{\mathbf{c}}^{\mathbf{L},+/-}$	In/Out-closeness	✓	In-vehicle travel time
Space-of-Service	$\mathbf{d}^{\mathbf{P},+/-}$	In/Out-degree	✗	–
	$\mathbf{b}^{\mathbf{P}}$	Betweenness	✗	–
	$\tilde{\mathbf{b}}^{\mathbf{P}}$	Betweenness	✓	Waiting time

used in the following section 7.2.3.

7.2.3 Independent variables: centrality indicators

Since the introduction of the "centrality" concept by Bavelas (1948), a variety of network centrality indicators have been proposed in the past decades. In principal, all these indicators are designed to capture distinct aspects of what it means to be "central" in a network for individual nodes. Based on this concept, this study employs several different centrality indicators for both space-of-infrastructure and space-of-service networks as the proxies of different properties of PTNs. The combination of different topological representations and centrality indicators enables a concise way to quantify a range of fundamental properties of PTNs. Moreover, some centrality indicators are computed in time-dependent weighted networks, which correspondingly reflect time-dependent characteristics of PTNs.

A summary of all the employed centrality indicators is shown in Table 7.1 and detailed descriptions are presented in the following subsections. General definitions of the selected centrality are first given, followed by their interpretation in different topological representations of PTNs. In addition, all the centrality indicators of the nodes are scaled by having the division over the sum for comparability and transferability reasons.

In/out-degree centrality

For unweighted directed networks, the *in/out-degree* centrality is an indicator that determines the importance of a node based on the number of links connected to it in an inbound/outbound manner. This indicator can be further extended by adding weights to the network as proposed by Barrat et al. (2003), which is coined by them as *strength*. We stick to the term "degree in weighted networks" in the remaining of this study for

the consistency with other indicators. The calculation of the in/out degree centrality in a weighted network can be based on the adjacency matrix A of it shown as follows:

$$\tilde{d}_i^+ = \sum_j w_{ji} A_{ji} \tag{7.1}$$

$$\tilde{d}_i^- = \sum_j w_{ij} A_{ij} \tag{7.2}$$

where $\tilde{d}_i^+$ and $\tilde{d}_i^-$ respectively denotes the in- and out- degree centrality of node i in a weighted network. w_{ij} denotes the value of weight of the corresponding link. When there is no weight considered , namely $w_{ij} = 1$, the indicators are degraded to the in- and out- degree centrality of node i in an unweighted network, denoted by d_i^+ and d_i^-.

- $\mathbf{d}^{\mathbf{L},+/-}$: In/out-degree centrality in the **unweighted** space-of-infrastructure network
  This indicator corresponds to the number of road or rail links that directly lead in or out of a given stop. It thus directly relates to the underlying physical infrastructure of PTNs.
- $\tilde{\mathbf{d}}^{\mathbf{L},+/-}$: In/out-degree centrality in the **weighted** space-of-infrastructure network
  Links are weighted by the time-dependent vehicle frequency between two adjacent stops with all the routes considered. Hence, this indicator quantifies the scheduled service intensity in terms of PT vehicle flows.
- $\mathbf{d}^{\mathbf{P},+/-}$: In/out-degree centrality in the **unweighted** space-of-service network
  This indicator measures the number of stops that can be reached without transfer for a given stop. It thus directly relates to the underlying service design of PTNs.

Betweenness centrality The *betweenness* centrality is a widely used indicator that was initially proposed by Freeman (1977) for social network studies. It quantifies the importance of a node in a network by measuring the proportion of the shortest paths between all node pairs in the network that pass through it. Assuming that flow travels through a network along the shortest path, nodes that lie on many shortest paths will undertake a high proportion of traffic, thus becoming more central in the network. In this sense, such a node might play a significant role in the passage of traffic through the network. The definition of the betweenness centrality is given as follows:

$$b_i = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{7.3}$$

where b_i denotes the betweenness centrality of node i. σ_{st} is the total number of shortest paths from node s to node t and $\sigma_{st}(i)$ is the number of those paths that pass through i.

Computing the betweenness centrality involves searching for all the shortest paths between node pairs. In this study, instead of leveraging on one single betweenness centrality indicator, the betweenness centrality indicators in both topological representations - i.e. the space-of-infrastructure and the space-of-service - are considered. The major advantage is that through their inclusion, we are able to directly estimate the contribution of each cost component: in-vehicle time (weighted space-of-infrastructure), number of transfers (unweighted space-of-service) and waiting time (weighted space-of-service). Consequently, their contributions to model prediction power are disentangled without pre-specifying any behavioral trade-offs in this process. An additional advantage is that the computational burden of the betweenness centrality is also greatly relieved in this way since the algorithm proposed by Brandes (2001) can be easily applied in our case. The betweenness centrality indicators in different topological networks are explained below.

- $\mathbf{b^L}$: Betweenness centrality in the **unweighted** space-of-infrastructure network
  The share of shortest paths that traverse a certain stop when path length is measured in terms of the number of stops traversed. Given some evidence (e.g., Guo, 2011), this indicator may coincide with how travelers choose their routes in complex PTNs using the map provided by agencies/operators as a mean to approximate travel time.

- $\tilde{\mathbf{b}}^{\mathbf{L}}$: Betweenness centrality in the **weighted** space-of-infrastructure network
  With the network weighted by the in-vehicle travel time, this indicator corresponds to the share of shortest paths in terms of on-board travel time that traverse the respective stop. Note that no regard is made to line configuration and thus the number of transfers induced.

- $\mathbf{b^P}$: Betweenness centrality in the **unweighted** space-of-service network
  This indicator relates to the interchange (hub) function of the respective stop. It therefore pertains to one of the most important and unique properties of PT systems, namely transfers.

- $\tilde{\mathbf{b}}^{\mathbf{P}}$: Betweenness centrality in the **weighted** space-of-service network
  The share of shortest paths measured in terms of the average waiting time that traverse a given stop. The path cost consists of the waiting time at the first stop for the route chosen and the waiting time at all subsequent transfer locations.

In/out-closeness centrality The intuition of the *closeness* centrality is that two nodes in a network are maximally close – in a topological sense – if they share a direct connection, whereas two nodes that are only tied indirectly through many intermediate nodes are topologically distant (Bavelas, 1950). Given this logic, the topological distance between two nodes can be defined as the number of links on the shortest path between them. Hence, a node becomes topologically central if it is able to interact with many network elements via only a few links, namely having a short average path

length. More formally, the closeness centrality of a node can be defined as the inverse of its average shortest path length (Beauchamp, 1965):

$$c_i = \frac{N-1}{\sum_{j \neq i} l_{ij}} \tag{7.4}$$

where c_i denotes the closeness centrality of node i. l_{ij} is the shortest path length, or topological distance, between nodes i and j. N is the number of nodes in the network. In directed networks, if we define that l_{ij} is the shortest path from node j to node i, then Equation 7.4 depicts the closeness centrality according to the shortest paths that are incoming to node i, which is defined as the *in-closeness* centrality. Similarly, the *out-closeness* centrality is based on the paths outgoing from node i, in which case we would instead sum over l_{ji} for $j = 1, ..., N$ in Equation 7.4. In weighted networks, the closeness centrality can be estimated by searching the shortest weighted path length between regions, where the weight of the path is determined by the sum of the link weights on that path.

- $\mathbf{c}^{\mathbf{L},+/-}$: In/out-closeness centrality in the **unweighted** space-of-infrastructure network
  This indicator quantifies the phenomenon that passengers originating from the topologically central stops can reach the others in the network with fewer intermediate ones.
- $\tilde{\mathbf{c}}^{\mathbf{L},+/-}$: In/out-closeness centrality in the **weighted** space-of-infrastructure network
  The weight is determined by the scheduled in-vehicle travel time, thus making the shortest path more related to the PT service.

The closeness centrality in the space-of-service network is not included in model development because it reflects a concept very similar to the one obtained through the degree centrality in the same space ($\mathbf{d}^{\mathbf{P},+/-}$), namely identifying the stops that are most reachable with the least number of transfers.

7.2.4 Dependent variable: passenger flow distribution

The time-dependent passenger flow distribution at PT stops is leveraged as the dependent variable, denoted by $\mathbf{q}$. Here we define the passenger flow at a stop in PTNs as the sum of inflow, outflow and throughflow at this stop during specified time slices. Specifically, inflow and outflow respectively represent the amount of passengers entering (boarding)/exiting (alighting) the PT system at a stop, while throughflow represents the amount of passengers that pass through a stop without leaving PT vehicles. This definition of the passenger flow sufficiently characterizes how intensively the stops are used across the network. In addition, the absolute passenger flows are converted into

relative terms, i.e. flow share, at each stop divided by the sum of all stop flows across the network during the respective time slices. we do not attempt to directly predict absolute flow values based on scaled centrality indicators because the same centrality value may correspond to different contexts for different networks and time periods. Instead, we examine whether the distribution of passenger flows is correlated with service properties by considering each stop and time-period as a single observation. Absolute flow values are resorted by multiplying flow shares by the total passenger flow in the network.

7.2.5 Model development

The model development is performed in two steps, with the first being an exploratory analysis among variables based on the Pearson correlation coefficient, and the second being building regression models. The objective of the first step is to find out (i) which independent variables (centrality indicators) have higher correlation with the dependent variable (passenger flow distribution), and thus can be incorporated into the models to be developed; (ii) the collinearity among independent variables. This is to ensure that variables that are mutually linearly correlated are not included in the models at the same time so that the developed models are as parsimonious as possible.

Following the exploratory analysis, we estimate regression models to capture the correlative relation between passenger flow distribution and network properties. Each observation in the dataset corresponds to the flow share associated with a given stop for a given time instance. Hence, the dataset contains (balanced) panel data, which are time-dependent and cross-sectional. Panel data regression models are thus applied. Let us denote the number of time periods for which each element (i.e. stop) i is observed as T_i. Panel data models are most useful when the outcome variable is expected to depend on explanatory variables which are not directly observable but correlated with the observed explanatory variables. If such omitted variables are time-invariable, panel data estimators allow to consistently estimate the effect of the observed explanatory variables. A general formulation of the panel data regression model with specific individual effects is presented below:

$$y_{it} = \alpha + \beta X_{it} + \mu_i + v_{it}, i = 1, ..., n, t = 1, ..., T_i \tag{7.5}$$

where μ_i represents the ith invariant time individual effect and $v_{it} \sim i.i.d(0, \theta_v^2)$ the disturbance. There are several different estimators (e.g., fixed effects, random effects, mixed effects, etc.) for panel data models based on different assumptions, of which more details can be found in relevant literature (e.g., Hsiao, 2007). In this study, the *random effects (RE)* model is applied in order to relieve the loss of degree of freedom during the estimation, as the number of units in our case is quite large (hundreds of PT stops). In RE models, the individual-specific effect is assumed to be a random

variable that is uncorrelated with the explanatory variables, i.e., $\text{Cov}(X_{it}, \mu_i) = 0$ and $\text{Cov}(X_{it}, v_{it}) = 0$ for all i and t. The model can be then formulated as:

$$y_{it} = \alpha + \beta X_{it} + \mu_{it}, i = 1, ..., n, t = 1, ..., T_i \tag{7.6}$$

where $\mu_{it} = \mu_i + v_{it}$ represents the error term that includes the ith invariant time individual effects μ_i and the disturbance v_{it}.

7.2.6 Model evaluation

Absolute passenger flows are used in the evaluation of the estimated models. These values, which are also time-dependent, are derived by multiplying the relative one (flow shares) that are obtained from the models by the total amount of flows in the network. Four evaluation measures, including the Mean Absolute Error (MAE), Weighted Mean Absolute Error (WMAE), Weighted Absolute Percentage Error (MAPE) and Weighted Mean Absolute Percentage Error (WMAPE). The motivation for taking into account the weighted measures - of which weights are determined by the magnitude of passenger flows at the corresponding stops - is that we want to reduce the bias caused by extreme error values at stops with low passenger flows. The applied measures are specified as below:

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n} \tag{7.7}$$

$$WMAE = \frac{\sum_{i=1}^{n} w_i \times |\hat{y}_i - y_i|}{\sum_{i=1}^{n} w_i} \tag{7.8}$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{7.9}$$

$$WMAPE = \frac{100\%}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \times \left| \frac{\hat{y}_i - y_i}{y_i} \right| \tag{7.10}$$

where $\hat{y}_i$ and y_i denote the predicted and actual passenger flows of stop i, respectively. w_i represents the weight of stop i, namely the flow share. n denotes the total number of observations for the evaluation data set.

7.3 Studied networks and experimental setup

7.3.1 Networks and data

The tram networks of two Dutch cities – The Hague and Amsterdam – were used for this investigation given the rich data availability of the Dutch PT systems. The

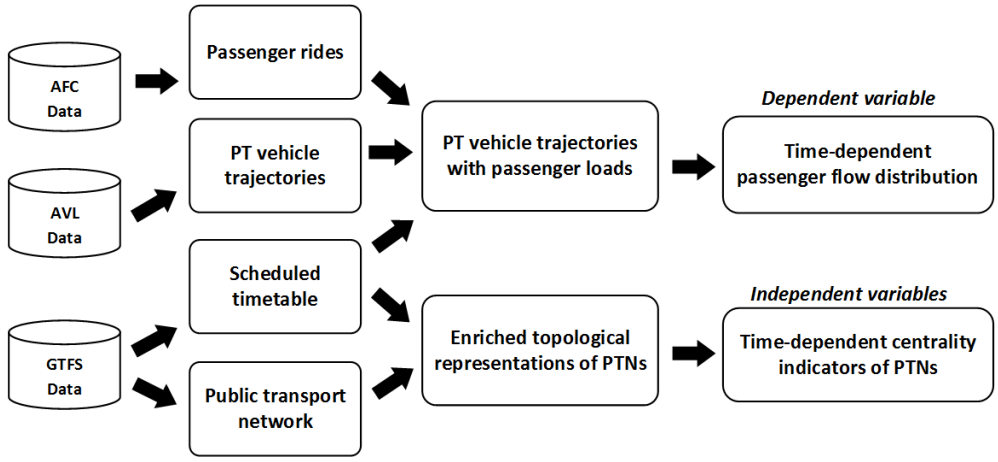


Figure 7.3: Workflow of the data preparation.

Table 7.2: Summary of the studied tram networks.

Quantity	The Hague	Amsterdam
Nodes	232	192
Directional Routes	28	24
Links in Space-of-Infrastructure	520	418
Links in Space-of-Service	8901	6122

data of the entire month of March, 2015 for The Hague, and that of the entire day of November 14th, 2017 for Amsterdam were leveraged. As Figure 7.3 shows, AFC, AVL, and GTFS data were used as major sources to generate networks as well as highly aggregated spatiotemporal data sets of dependent and independent variables. The passenger flow flow distribution (dependent variable) were obtained from the PT vehicle trajectories with passenger loads (Luo et al., 2018). A summary of the basic properties of the two networks, including the number of nodes, (directional) routes, links in space-of-infrastructure and space-of-service networks, is presented in Table 7.2.

7.3.2 Experimental setup

For the experimental setup, we selected 20 working days with normal demand patterns (out of one month) for The Hague. 15 working days were further randomly selected for the model development, with the data aggregated on an hourly basis from 6 am to 12 am (18 time slices). The rest five-day data set of The Hague and the one-day data set of Amsterdam were utilized for the model evaluation.

7.4 Results and discussion

The results of the exploratory analysis on the two employed networks are first shown in section 7.4.1. Section 7.4.2 then presents the results of model estimation, followed by the model evaluation in section 7.4.3.

7.4.1 Exploratory analysis

To gain more intuition about the spatial distribution of passenger flow and centrality indicators in the studied networks, the visualizations of them for the weekday morning peak (7 am - 8 am) are performed and presented in Figure 7.4 (The Hague) and Figure 7.5 (Amsterdam). Both size and color are used to make the distinction in magnitude remarkable. Out-degree and out-closeness are omitted as they display the same pattern as their counterparts. Through the visualizations, it can be seen that considerable amount of passenger flows are loaded in the central area of both networks, though it is also observable that some corridors used by commuters also undertake a significant amount of flows, such as the one from center to the east in The Hague, and two horizontal corridors in Amsterdam with one on the middle of west and the other on the top of east. We can further notice that the in-degree centrality in the weighted space-of-infrastructure ($\tilde{\mathbf{d}}^{\mathbf{L},+}$) and the betweenness centrality in both unweighted ($\mathbf{b}^{\mathbf{L}}$) and weighted ($\tilde{\mathbf{b}}^{\mathbf{L}}$) space-of-infrastructure mostly match the flow distribution pattern with clear distinctions among nodes across the networks. Some indicators, including the in-degree in the unweighted space-of-infrastructure ($\mathbf{d}^{\mathbf{L},+}$) and the in-closeness in both unweighted ($\mathbf{c}^{\mathbf{L},+}$) and weighted ($\tilde{\mathbf{c}}^{\mathbf{L},+}$) space-of-infrastructure, show rather plain patterns. Besides, the betweenness in the space-of-service ($\mathbf{b}^{\mathbf{P}}$ and $\tilde{\mathbf{b}}^{\mathbf{P}}$) makes the transfer locations in the networks really stand out.

The temporal variance in the dependent and independent variables for both networks is further displayed through the distribution plots in Figure 7.6. Four different time slices are selected for each variable to demonstrate the distinction between peak (07:00 - 08:00 and 17:00 - 18:00) and non-peak (11:00 - 12:00 and 22:00 - 23:00) periods. For the dependent variable, i.e. share of passengers flow, a few nodes are traversed by a large proportion of the flows while the rest of them only share a small proportion. This pattern is persistent over all time periods and is observed for both networks. As for the independent variables, significant differences across the four time periods can be observed for $\tilde{\mathbf{d}}^{\mathbf{L},+}$ since it largely depends on the planned service frequency which varies over the day. Differences also hold for $\tilde{\mathbf{b}}^{\mathbf{L}}$, albeit to a lesser extent, due to different traffic conditions.

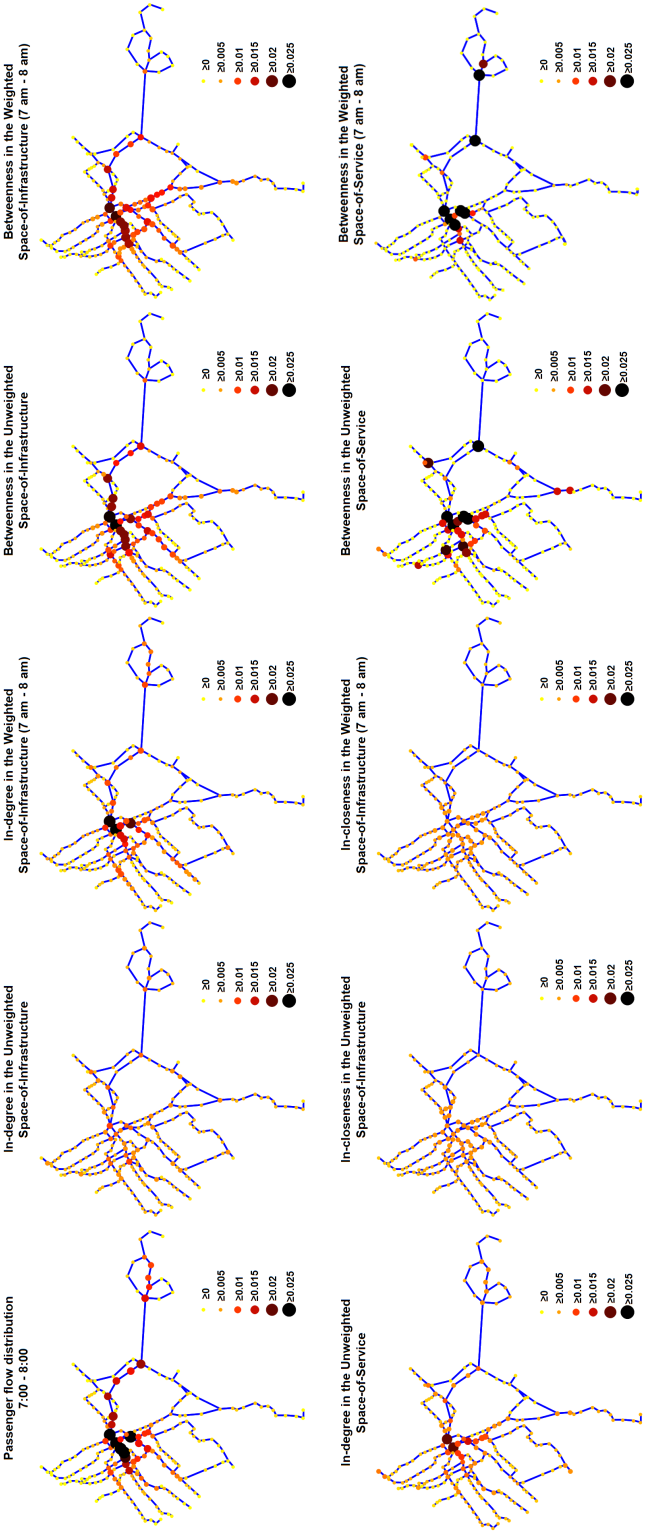


Figure 7.4: Visualization of the passenger flow distribution and the employed centrality indicators for the weekday morning peak (7am - 8 am) of the tram network of The Hague.

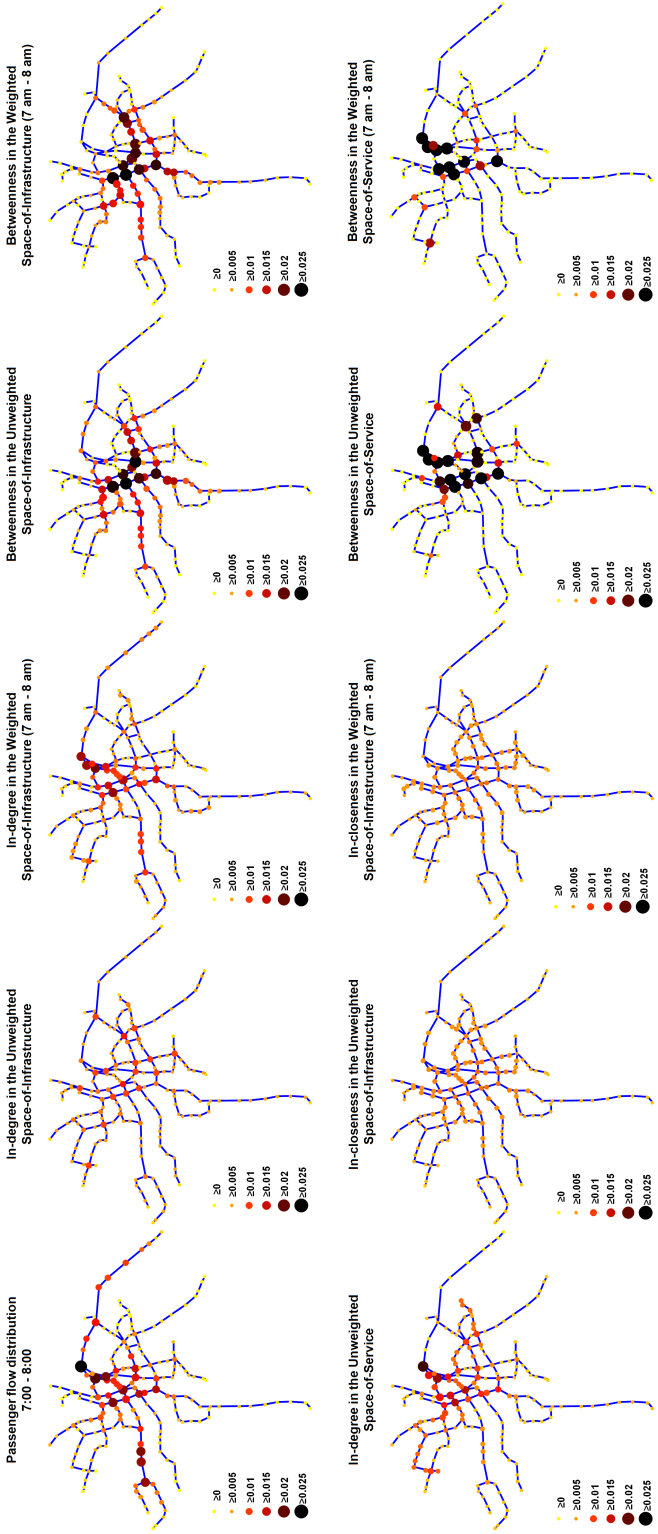


Figure 7.5: Visualization of the passenger flow distribution and the employed centrality indicators for the weekday morning peak (7am - 8 am) of the tram network of Amsterdam.

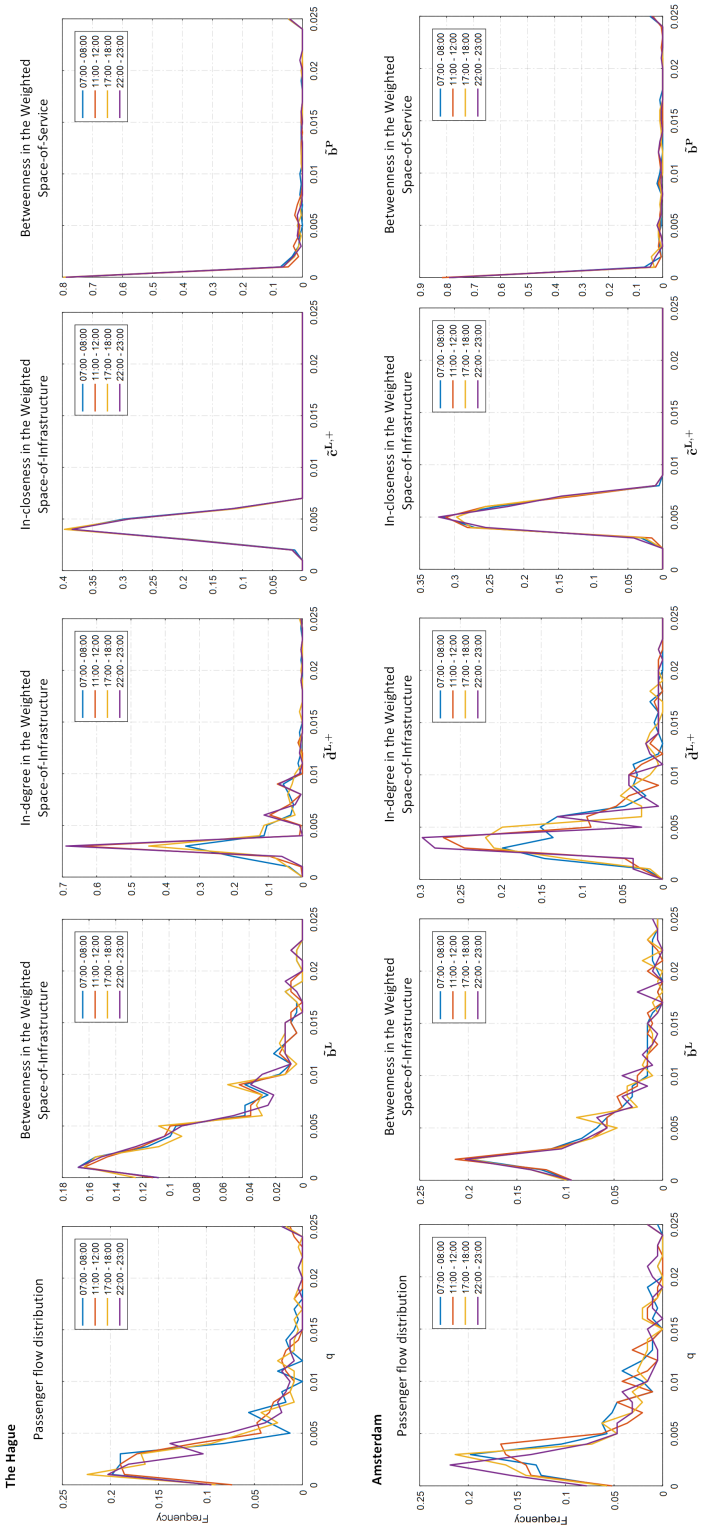


Figure 7.6: Illustration of the temporal variation of the distributions of dependent and independent variables for both The Hague and Amsterdam tram networks. Four different time slices are selected to display mainly the difference between peak (07:00 - 08:00 and 17:00 - 18:00) and non-peak periods (11:00 - 12:00 and 22:00 - 23:00).

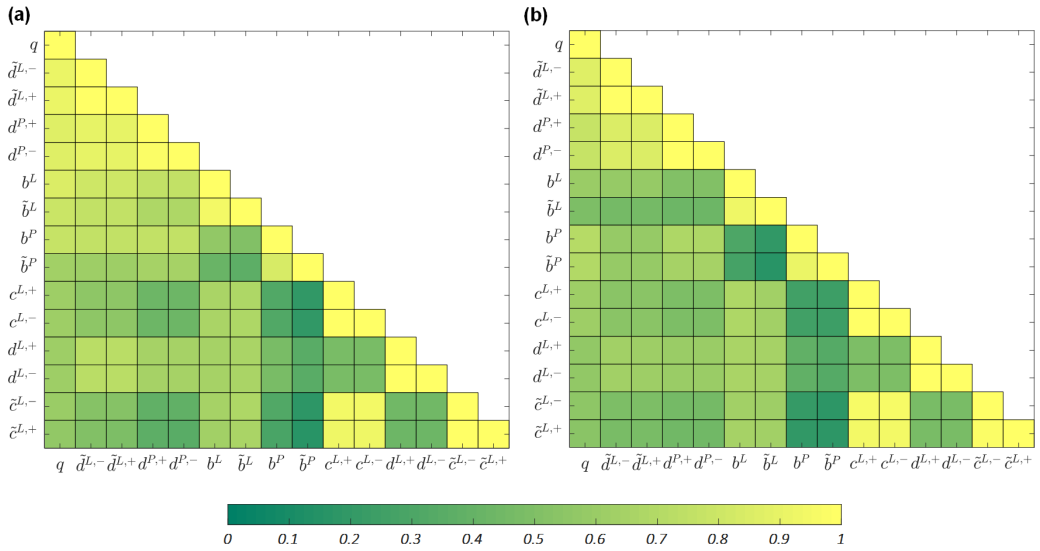


Figure 7.7: Illustration of the Pearson correlation coefficient matrices among different variables. (a) The Hague; (b) Amsterdam.

We further visualize the correlation coefficient matrices among all the variables for both networks as shown in Figure 7.7. With the passenger flow distribution $\mathbf{q}$ placed at the first place on both x and y axis, the sequence of the centrality indicators is arranged in a descending manner based on the correlation between them and $\mathbf{q}$. The sequence for both diagrams is determined by the case of The Hague for the sake of model development.

According to Figure 7.7, the in/out- degree centrality indicators in the weighted space-of-infrastructure network ($\tilde{\mathbf{d}}^{\mathbf{L},+/-}$) show the highest positive correlation with $\mathbf{q}$ in both The Hague and Amsterdam tram systems. This is consistent with the visual patterns from Figure 7.4 and Figure 7.5. It is also intuitive to interpret because the amount of passengers that is moved in the network depends on the PT vehicle flows. The following indicators are the in/out- degree centrality in the unweighted space-of-service networks ($\mathbf{d}^{\mathbf{P},+/-}$). Note that these two indicators also show high correlation with the previous ones. They are thus not considered when the in/out- degree centrality in the weighted space-of-infrastructure network are used in the model development.

The group of degree centrality indicators are followed by the betweenness ones. Note that in the case of The Hague (Figure 7.7a), the values of betweenness centrality in both of the unweighted and weighted space-of-infrastructure networks ($\mathbf{b}^{\mathbf{L}}$ and $\tilde{\mathbf{b}}^{\mathbf{L}}$) are higher than those in the unweighted and weighted space-of-service networks ($\mathbf{b}^{\mathbf{P}}$ and $\tilde{\mathbf{b}}^{\mathbf{P}}$). This, nevertheless, is opposite in the Amsterdam system (Figure 7.7b). In fact, the betweenness centrality in the space-of-infrastructure does not seem to be a good proxy to the passenger flow distribution for the Amsterdam tram network. It performs even worse than the closeness centrality in the space-of-infrastructure. The remaining centrality indicators are presented in the end as they do not show significantly high correlation with $\mathbf{q}$.

Table 7.3: Estimation results of the selected models.

Independent	Model 1 (OLS)			Model 2 (RE)			Model 3 (RE)			
Variables	Coef.	Std.Err	t-stat	Coef.	Rob.Std.Err	z-stat	Coef.	Rob.Std.Err	z-stat	VIF
CONST	-0.0003	0.0003	-1.3670	0.0022 ***	0.0004	5.8880	-0.0029 ***	0.0005	-6.1798	-
b^L	1.0799 ***	0.0397	27.2241	-	-	-	0.5951 ***	0.0699	8.5133	3.7495
$\tilde{d}^{L,+}$	-	-	-	0.4849 ***	0.1168	4.1511	0.1135 ***	0.0554	2.0491	5.1796
$c^{L,+}$	-	-	-	-	-	-	0.8419 **	0.1244	6.7693	1.8418
b^P	-	-	-	-	-	-	0.1140 ***	0.0089	12.7630	2.5844
Num. Obs.		232			4176			4176		
R^2		0.7632			0.85723			0.89954		
Adj R^2		0.7621			0.85720			0.89944		

*** $p < 0.01$ ** $p < 0.05$

7.4.2 Model estimation

The model estimation was performed using MATLAB, with the RE models estimated using the panel data toolbox developed by Álvarez et al. (2017). Note that the robust standard error estimation of the RE models was computed when accounting for heteroscedasticity. Moreover, the variance inflation factor (VIF), which quantifies the severity of collinearity in a regression model, was also computed for the parameters of Model 3 which includes several independent variables.

Three model estimations based on the training dataset from The Hague are presented and discussed in this section, with the detailed results displayed in Table 7.3. Note that Model 1 is estimated as an ordinary least squares model. This is because of the fact that there is no temporal variance in the only independent variable ($\mathbf{b^L}$), and the temporal dimension of the independent variable ($\mathbf{q}$) is correspondingly also canceled by summing up the flows over all periods. This model indicates to what extent it is possible to approximate the global passenger flow distribution using solely topological information without embedding time-dependent service attributes. The other two models, Model 2 and Model 3, are estimated using RE models as explained in section 7.2.5 because both of them incorporate independent variables pertaining to frequency which is time-dependent. The third model has the highest prediction power also when accounting for the number of parameters included (Adjusted R^2). It includes four centrality indicators: betweenness centrality in the unweighted space-of-infrastructure ($\mathbf{b^L}$), in-degree centrality in the weighted space-of-infrastructure ($\tilde{\mathbf{d}}^{\mathbf{L},+}$), betweenness centrality in the unweighted space-of-service ($\mathbf{b^P}$), and in-closeness centrality in the unweighted space-of-infrastructure ($\mathbf{c}^{\mathbf{L},+}$). The VIF values confirm that all the incorporated independent variables in Model 3 do not exercise significant collinearity since all the values are lower than 10 (Marquardt, 1980).

7.4.3 Model evaluation

The estimated models are evaluated for the tram networks of The Hague (evalution dataset) and Amsterdam. The results are summarized in Table 7.4. Note that the evaluation is performed based on the absolute flows obtained by multiplying the pre-

Table 7.4: Results of the evaluation metrics for the selected models.

Evaluation	The Hague			Amsterdam		
Metrics	Model 1	Model 2	Model 3	Model 1	Model 2	Model 3
MAE (pax)	184	260	128	335	305	240
WMAE (pax)	520	841	283	804	715	452
MAPE (%)	77.6	248.3	70.9	71.4	155.7	68.8
WMAPE (%)	42.0	58.7	29.1	55.6	50.6	39.8

dicted relative flow shares by the total amount of flows in the network. Unsurprisingly, Model 3 largely outperforms Model 1 and Model 2 regardless of the metric used. This suggests that models based on a single centrality indicator that does not incorporate information also from the space-of-service are not able to well capture the correlation. In addition, the discrepancy between weighted and unweighted metrics is striking, implying that significant predictive errors occur to stops with relatively low flows.

Further, we plot the actual versus predicted flows for Model 3 in Figure 7.8a and Figure 7.8d. It can be observed that Model 3 is indeed well-able to predict passenger flows in both networks. It is also evident that the model performs particularly well when the predictions are made for the same network for which the data has been trained (The Hague).

The spatial distribution of evaluation errors in both absolute and relative terms are also visualized and presented in Figure 7.8. Both negative and positive values are considered in the visualizations, corresponding to underestimations (blue) and overestimations (red), respectively. Plots in Figure 7.8b and Figure 7.8e show absolute error terms, while those in Figure 7.8c and Figure 7.8f show relative error terms. In the case of The Hague, it can be observed from Figure 7.8c that large relative over- or under-estimations occur at stops located further away from the center. However, these relatively large errors in relative terms are small in absolute terms as can be seen in Figure 7.8b. In absolute terms, flows at stops in the core of the network tend to be underestimated, while flows along corridors that offer cycles between main parts of the network such as along cross-radial lines are mostly overestimated. Similar overall patterns are observed in the case of Amsterdam, albeit with larger absolute deviations resulting from larger overall demand levels. Hence, flows in the very central core of the network around the central station and the key tourist attractions are underestimated while the flows along the two half-circular infrastructure is overestimated (in both relative and absolute terms for both cases).

7.5 Conclusion

This chapter presents a pioneering investigation into the relation between passenger flow distribution and network properties in PT systems. Differing from the traditional approach that consists of demand estimation and assignment, this study is performed in

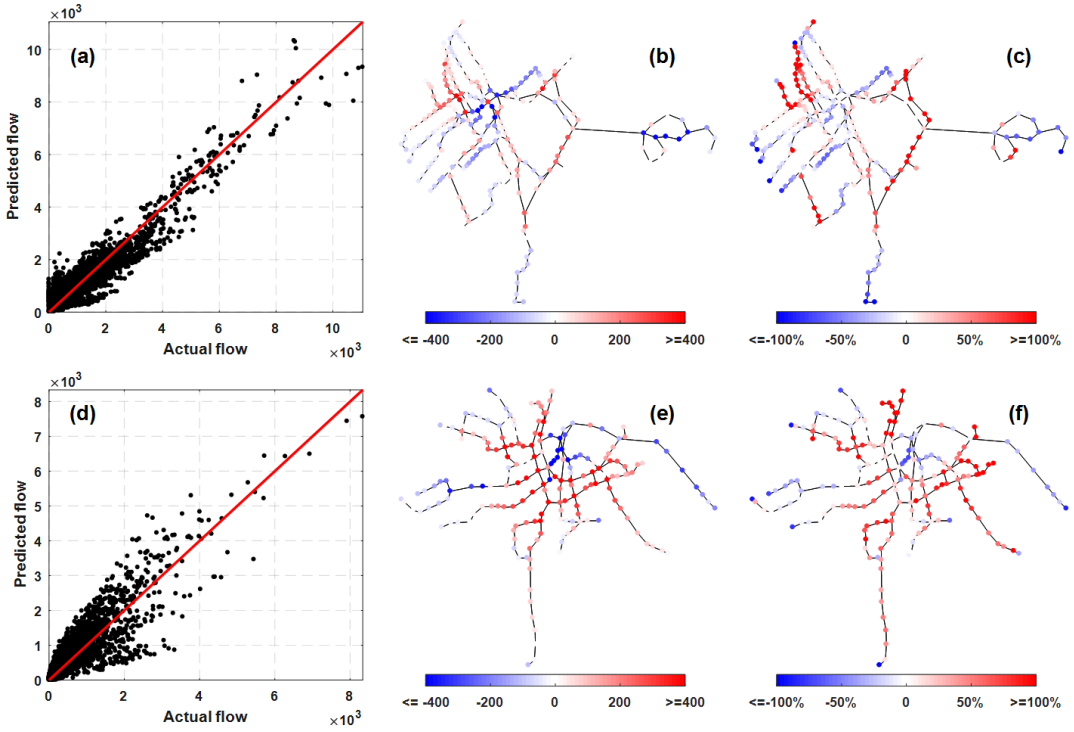


Figure 7.8: Illustrations of the evaluation errors for Model 3. (a) Actual flow vs predicted flow plot for The Hague; (b) Spatial distribution of the absolute errors for The Hague; (c) Spatial distribution of the relative errors for The Hague; (d) Actual flow vs predicted flow plot for Amsterdam; (b) Spatial distribution of the absolute errors for Amsterdam; (c) Spatial distribution of the relative errors for Amsterdam.

a reverse engineering fashion by directly examining the relation between the observed flow distribution and network properties that are quantified by centrality indicators in various topological representations of PTNs. This research capitalizes on the capability to measure PT systems using passively collected PT data (e.g., AFC, AVL and GTFS). In addition, concepts and methods adopted from complex network science, including the topological representation of PT infrastructure and service networks and centrality indicators, also play a key role in a sense that the combination of them provides a systematic and concise way to quantify the network properties of PT systems. All the employed centrality indicators are also interpreted in the context of PT systems, which enriches the application of complex network science in the transport research.

The major conclusion drawn from the case study on the tram networks from The Hague and Amsterdam is that the selected network properties can indeed be used to approximate the global passenger flow distribution across the network to a reasonable extent of accuracy using solely regression models. This however does not imply causality as it is likely that supply provision has been designed to correspond to demand patterns and therefore the relation reflects the interplay between demand and supply distributions. Based on the evidence presented in this chapter, several research directions can be further explored in the future. First, more real-world PT systems can be employed in

order to further validate the finding. Second, the proposed approach can be instrumental in a range of PT applications. This includes conducting full-scan evaluations of the impact of planned disruption on the redistribution of passenger flows throughout the network, which can serve as a good complement to the prevailing tools, i.e., simulation models, at a much lower computational cost and with fewer assumptions. Third, the extent to which PT supply is well designed to reflect passenger flow distribution can be considered as a network performance metric for monitoring system performance over time as well as comparing alternative networks.

Future research can explore the following directions. First, more real-world PT systems shall be researched to further validate the conclusion of this study. Second, the proposed method itself can be extended to a range of PT applications. This includes conducting full-scan evaluations of the impact of planned disruption on the redistribution of passenger flows throughout the network, which can serve as a good complement to the prevailing tools, i.e., simulation models, at a much lower computational cost and with fewer assumptions. Third, the extent to which PT supply is well designed to reflect passenger flow distribution can be considered as a network performance metric for monitoring system performance over time as well as comparing alternative networks.

Chapter 8

Conclusions, Implications and Future Research

The research presented in this thesis has been centered around the primary objective of developing methods and models for translating high-volume data from a variety of sources into new knowledge and insights which can be used to improve public transport planning and operations. Five specific research questions have been investigated in the previous chapters. In this final one, we draw conclusions in section 8.1 and implications for practice in section 8.2, respectively. We further discuss research directions that are worthwhile exploring in the future in section 8.3.

8.1 Conclusions

This thesis has been devoted to developing methods and models for translating high-volume data from a variety of sources into novel knowledge and insights that can be used to improve public transport planning and operations. To conclude the thesis, we now provide answers to all the research questions raised in **Chapter 1**.

RQ1: How can we generate more information-rich profiles of PT vehicles containing positions and onboard occupancy based on prevalent PT data sources? (**Chapter 3**)

To solve this problem, we first specifically identified the data issues related to each and the combination of different sources, including AFC, AVL and GTFS. We then proposed a method for systematically addressing these issues and further constructing profiles of PT vehicles based on different data sources. The method consists of four steps through which raw data from different sources are processed and integrated. These steps are (i) data pre-processing; (ii) matching trips in GTFS and AVL; (iii) matching passenger rides to vehicle trajectories; (iv) improving vehicle trajectories. The resulting vehicle profiles contain detailed information of both vehicles' scheduled and actual trajectories and passenger loads. We demonstrated the proposed method using the data from the PT system of The Hague, the Netherlands. The profiles were further visualized using space-time seat occupancy diagrams, which provides operators with a compact and powerful reference to intuitively examine the onboard crowding patterns over time and space. This visualization technique can help operators in timetable optimization, network and fleet scheduling, and sub-route service designing.

RQ2: How can we reduce the high dimensionality of passenger flows for large-scale analysis and modeling? (**Chapter 4**)

We applied principal component analysis (PCA) to address this challenge. We first showed how the matrix of multivariate time series of passenger flows can be constructed, and then specified how such high-dimensional flow matrix can be transformed using PCA. To demonstrate the method, we employed the metro system in Shenzhen, China for the case study, in which entry and exit flows of all the stations – which were derived from a one-month AFC data set with both tap-in and tap-out records – were considered simultaneously. The result showed that a great amount of variance contained in the original data can be effectively retained in lower-dimensional subspaces composed of top few principal components. We further analyzed the features of such low dimensionality, with detailed investigations into the principal components and temporal stability of the flow structure. This study contributes to the understanding of large-scale PT flow dynamics. It also paves the way for integrating PCA into large-scale passenger flow modeling and predictions.

RQ3: How can we construct zone-to-zone OD matrices for PT systems using data-driven techniques? (**Chapter 5**)

We proposed a novel data-driven method based on the k-means algorithm. The novelty pertains to heuristically determining the number of clusters based on both spatial and

passenger flow patterns. The proposed method consists of four steps. First, the best clustering of each K is constructed by running the k-means algorithm a number of times with different initial centers. The one that results in the minimum SSE, a measure for the variance of clusters, is selected. The following two steps pertain to computing two metrics based on distance and passenger flow respectively based on both intra-cluster and inter-cluster components. The last step combines the two metrics to determine the optimal number of cluster following the criterion adopted. Differing from the traditional way of grouping stops based on predefined traffic analysis zones, our proposed method provides a data-driven perspective for solving such problems using passenger flows that can be directly observed rather than their proxies. The method was demonstrated by a case study of the PT system of The Hague, the Netherlands based on a data set of one-month stop-to-stop OD flows derived from the local AFC data. We adopted the criterion that the ratio of average intra-cluster flow to average inter-cluster flow should be maximized while maintaining the spatial compactness of all clusters. This type of clustering for generating zone-to-zone OD matrices is particularly suitable for urban areas with high density of PT stops, because it allows for integrating travelers' (origin and destination) stop and route choices into subsequent modeling studies. This analysis process can be applied using other spatial and flow metrics of interest, depending on the application, case study characteristics and data availability. The temporal variability analysis shows that the variance in passenger flow over time does not have a significant influence on the final determination of number of clusters when using the proposed method, which implies that this method is robust and can be potentially adopted for both short-term and long-term PT related research.

RQ4: How can we analyze the accessibility of public transport networks in a more transferable and efficient way? (**Chapter 6**)

We addressed this issue by proposing a new method which integrates network science and PT accessibility analysis. The average travel impedance was used in this study to measure PT accessibility. We proposed an innovative weighted graph representation of PTNs that explicitly incorporates travel costs according to the planned services contained in GTFS data, which allowed for efficient computation of minimal generalized travel costs between stop pairs. Such cost is comprised of initial and transfer waiting times, in-vehicle travel times and time-equivalent transfer penalty costs. To demonstrate the high transferability and efficiency of the proposed method, we applied it to assess worldwide tram networks' accessibility. Such latitudinal comparative assessments can provide additional insights into PTN design, benchmark and planning, but are still scarce in the current literature due to the requirements imposed by existing methods that are heavily reliant on GIS. The case study has led to following new insights. According to the comparison between the benchmark and proposed GTC-based metrics, the main conclusion is that the spatial disparity in PT accessibility can be higher when planned service properties (i.e., travel times, initial and transfer waiting times, and transfer penalties) are considered than when they are not. In addition, the subsequent investigation shows that larger networks in terms of the spatial scale indeed

exhibit higher median total travel impedance. However, this is primarily attributed to the component of in-vehicle travel times which is intrinsically positively correlated to the spatial scale of networks. By excluding in-vehicle travel times, we can see that the majority of the networks exhibit a median cost of around 15 minutes regarding waiting and transfers, with the largest and smallest being close to 20 minutes (Budapest) and lower than 10 minutes (Toronto), respectively. Among all the studied networks, Milan shows the largest variance in this specific cost component.

RQ5: Can passenger flows be estimated solely based on the network properties of PT systems? (**Chapter 7**)

We conducted a pioneering empirical study endeavoring to obtain an answer to the proposed question. The study is performed in a reverse-engineering fashion by directly examining the relation between observed flow distributions and network properties that are quantified by centrality indicators in various topological representations of PTNs. It capitalizes on the capability to measure PT systems using passively collected PT data (e.g., AFC, AVL and GTFS). In addition, concepts and methods adopted from complex network science, including the topological representation of PT infrastructure and service networks and centrality indicators, allows for providing a systematic and concise way to quantify the network properties of PT systems. All the employed centrality indicators are also interpreted in the context of PT systems, which enriches the application of complex network science in the transport research. The major conclusion drawn from the case study on the tram networks from The Hague and Amsterdam is that the selected network properties can indeed be used to approximate the global passenger flow distribution across the network to a reasonable extent of accuracy using solely regression models. However, it should be noted that this conclusion does not imply causality. This is because it is likely that supply provision has been designed to correspond to demand patterns and therefore the reflects the interplay between demand and supply distributions. This finding can lead to the development of (i) parsimonious data-driven PT assignment models and (ii) global metrics for overseeing large-scale flow dynamics and the gap between flows and service supplies.

8.2 Implications for practice

Over the past years, it has become increasingly common for operators worldwide to collect massive amount of data from their PT systems. The costs for collecting and storing such data on a daily basis are not cheap, but many operators are still struggling to take advantage of these data for improving their operations due to the lack of insights into the data themselves. Moreover, the increasing pressure on data privacy issues, particularly in Europe given the latest General Data Protection Regulation (GDPR), has made operators even more cautious and reluctant to make proactive steps in exploiting their data resource. Therefore, practitioners need more specific evidence, guidance and visions from academia to further unlock the potential of data.

This thesis can hence serve as a timely reference to fulfill such need of PT planners and operators. Multiple implications for practice can be derived from this thesis's research outcomes, which are summarized as follows.

First, combining different data sources can significantly improve operators' capability in monitoring their systems' performance. Our research has demonstrated that multiple data sources, including AFC, AVL and GTFS, can be integrated to produce highly information-rich profiles (**Chapter 3**). These profiles allow for more in-depth analyses and visualizations of vehicle bunching and passenger crowding patterns, thus are quite instrumental in enhancing planning and operations. In addition, the identified data issues that are common across different systems can significantly benefit operators in terms of addressing their own data issues and upgrading data collection systems.

Second, new tools that can developed based on novel data-driven methods would help operators better achieve informed decision making in planning and operations. In this thesis, we have demonstrated new data-driven methods for zoning (**Chapter 5**), assessing accessibility (**Chapter 6**) and estimating passenger flows (**Chapter 7**). These potential tools will by no means completely replace existing ones, such as simulation models and GIS-based software. On the contrary, they will nicely serve as effective complementary approaches given their high applicability and computational efficiency.

Third, new metrics for measuring network-wide demand and supply dynamics can help operators more comprehensively evaluate their systems' performance. Our research has shown that dynamics of large-scale passenger flows can be well retained in low-dimensional spaces (**Chapter 4**). Furthermore, network-wide passenger flows can be approximated by the properties of service network (**Chapter 7**). These outcomes can thus be applied to design parsimonious global metrics for overseeing large-scale flow dynamics and the gap between flows and service supplies.

Last but not least, the data requirements from conducting research and building data-driven applications should already be well considered in the designing phase of data collection systems. This perhaps can be the most significant implication for practice as this type of requirements was largely overlooked by the current generation of PT systems. More intelligent and thorough designs of data collection systems would save a great deal of research efforts. For instance, making different data systems well inter-connected through common keys should be a fundamental standard. Besides, the systems should be able to log passengers' origin and destination stops more precisely, which would largely reduce the effort needed for developing inference algorithms. Real-time data transmission would also be much desired given the increasing demand for real-time applications. These measures are not difficult to realize given current technologies, yet will tremendously increase the data usability for both researchers and practitioners. In addition, the concept of privacy by design should also be applied to all the data systems. It should never become an excuse for preventing analyst from moving forward in the era of digitization.

8.3 Recommendations for future research

Besides specific future research directions that have already been discussed at the end of each previous chapters, we now provide the following ones that would be able to guide scholars on a higher level in the context of data-driven PT research.

The first direction is to develop **real-time predictive analytics**. According to Figure 2.4, real-time functions play an important role in assisting PT operations. Among them, real-time (large-scale) predictive analytics are even more significant, which however have not been extensively researched. A recent example was presented by Noursalehi et al. (2018) regarding the real-time demand prediction. For future research, short-term predictions on OD flows and onboard occupancy of PT vehicles would be much desired.

Second, more research can be conducted to unravel **spatiotemporal patterns of demand and supply of PT systems** using advanced network science and machine learning techniques, such as Sun et al. (2013); Sun & Axhausen (2016). These research will lead to more insights into complicated dynamics of PT systems, thus paving the road for developing more advanced tools for planning and operations.

Third, it is intriguing to study the **synergy between PT and Mobility as a Service (MaaS)** given the rapid rise of the later worldwide. Our society is facing bigger mobility challenges, but we are at the same time embracing new mobility providers, such as Uber, Via, Lyft. Whether they can co-exist with the existing PT and together offer a better solution to mobility challenges is still a open but very value question to answer by future research.

Bibliography

Akbarzadeh, M., S. Memarmontazerin, S. Derrible, S. F. Salehi Reihani (2019) The role of travel demand and network centrality on the connectivity and resilience of an urban street system, *Transportation*, 46(4), pp. 1–15.

Alfred Chu, K., R. Chapleau (2008) Enriching Archived Smart Card Transaction Data for Transit Demand Modeling, *Transportation Research Record: Journal of the Transportation Research Board*, 2063, pp. 63–72.

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

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: Journal of the Transportation Research Board*, 2535, pp. 88–96.

Altshuler, Y., R. Puzis, Y. Elovici, S. Bekhor, A. Pentland (2011) Augmented betweenness centrality for mobility prediction in transportation networks, in: *Finding Patterns of Human Behaviors in Network and Mobility Data (NEMO)*.

Álvarez, I. C., J. Barbero, J. L. Zofío (2017) A Panel Data Toolbox for MATLAB, *Journal of Statistical Software, Articles*, 76(6), pp. 1–27.

Anwar, A., A. Odoni, N. Toh (2016) BusViz: Big Data for Bus Fleets, *Transportation Research Record: Journal of the Transportation Research Board*, 2544, pp. 102–109.

Bagherian, M., O. Cats, N. van Oort, M. Hickman (2016) Measuring passenger travel time reliability using smart card data, in: *TRISTAN IX: Triennial Symposium on Transportation Analysis, Oranjestad, Aruba*.

Barabási, A. L., R. Albert (1999) Emergence of scaling in random networks, *Science*, 286(5439), pp. 509–512.

Barrat, A., M. Barthelemy, R. Pastor-Satorras, A. Vespignani (2003) The architecture of complex weighted networks, *Proceedings of the National Academy of Sciences of the United States of America*, 101(11), pp. 3747–3752.

Barry, J., R. Freimer, H. Slavin (2009) Use of Entry-Only Automatic Fare Collection Data to Estimate Linked Transit Trips in New York City, *Transportation Research Record: Journal of the Transportation Research Board*, 2112, pp. 53–61.

Barry, J., R. Newhouser, A. Rahbee, S. Sayeda (2002) Origin and Destination Estimation in New York City with Automated Fare System Data, *Transportation Research Record: Journal of the Transportation Research Board*, 1817, pp. 183–187.

Batty, M. (2009) Accessibility: In Search of a Unified Theory, *Environment and Planning B: Planning and Design*, 36(2), pp. 191–194.

Bavelas, A. (1948) A mathematical model for group structures, *Human Organization*, 7(3), pp. 16–30.

Bavelas, A. (1950) Communication Patterns in Task-Oriented Groups, *The Journal of the Acoustical Society of America*, 22(6), pp. 725–730.

Beauchamp, M. A. (1965) An improved index of centrality, *Behavioral science*, 10(2), pp. 161–163.

Berche, B., C. Von Ferber, T. Holovatch, Y. Holovatch (2009) Resilience of public transport networks against attacks, *European Physical Journal B*, 71(1), pp. 125–137.

Brandes, U. (2001) A faster algorithm for betweenness centrality, *The Journal of Mathematical Sociology*, 25(2), pp. 163–177.

Cathey, F. W., D. J. Dailey (2003) A prescription for transit arrival/departure prediction using automatic vehicle location data, *Transportation Research Part C: Emerging Technologies*, 11(3-4), pp. 241–264.

Cats, O. (2016) The robustness value of public transport development plans, *Journal of Transport Geography*, 51, pp. 236–246.

Cats, O. (2017) Topological evolution of a metropolitan rail transport network: The case of Stockholm, *Journal of Transport Geography*, 62, pp. 172–183.

Cats, O., E. Jenelius (2014) Dynamic Vulnerability Analysis of Public Transport Networks: Mitigation Effects of Real-Time Information, *Networks and Spatial Economics*, 14(3-4), pp. 435–463.

Cats, O., G. J. Koppenol, M. Warnier (2017) Robustness assessment of link capacity reduction for complex networks: Application for public transport systems, *Reliability Engineering and System Safety*, 167, pp. 544–553.

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.

Cats, O., M. Yap, N. van Oort (2016) Exposing the role of exposure: Public transport network risk analysis, *Transportation Research Part A: Policy and Practice*, 88, pp. 1–14.

Ceder, A. (2015) *Public Transit Planning and Operation : Modeling, Practice and Behavior, Second Edition*, CRC Press.

Cepeda, M., R. Cominetti, M. Florian (2006) A frequency-based assignment model for congested transit networks with strict capacity constraints: characterization and computation of equilibria, *Transportation Research Part B: Methodological*, 40(6), pp. 437–459.

Chu, K. K. A., R. Chapleau (2010) Augmenting Transit Trip Characterization and Travel Behavior Comprehension, *Transportation Research Record: Journal of the Transportation Research Board*, 2183, pp. 29–40.

Currie, G. (2010) Quantifying spatial gaps in public transport supply based on social needs, *Journal of Transport Geography*, 18(1), pp. 31–41.

Currie, G. (2018) Lies, Damned Lies, AVs, Shared Mobility, and Urban Transit Futures, *Journal of Public Transportation*, 21(1), pp. 19–30.

de Regt, R., C. von Ferber, Y. Holovatch, M. Lebovka (2019) Public transportation in Great Britain viewed as a complex network, *Transportmetrica A: Transport Science*, 15(2), pp. 722–748.

Derrible, S. (2012) Network centrality of metro systems, *PLoS ONE*, 7(7), p. e40575.

Derrible, S., C. Kennedy (2009) Network Analysis of World Subway Systems Using Updated Graph Theory, *Transportation Research Record: Journal of the Transportation Research Board*, 2112, pp. 17–25.

Derrible, S., C. Kennedy (2010a) Characterizing metro networks: State, form, and structure, *Transportation*, 37(2), pp. 275–297.

Derrible, S., C. Kennedy (2010b) The complexity and robustness of metro networks, *Physica A: Statistical Mechanics and its Applications*, 389(17), pp. 3678–3691.

Derrible, S., C. Kennedy (2011) Applications of graph theory and network science to transit network design, *Transport Reviews*, 31(4), pp. 495–519.

Devillaine, F., M. Munizaga, M. Trépanier (2012) Detection of Activities of Public Transport Users by Analyzing Smart Card Data, *Transportation Research Record: Journal of the Transportation Research Board*, 2276, pp. 48–55.

Djukic, T., J. Van Lint, S. Hoogendoorn (2012) Application of principal component analysis to predict dynamic origin-destination matrices, *Transportation Research Record: Journal of the Transportation Research Board*, 2283(1), pp. 81–89.

Dupuy, G. (2013) Network geometry and the urban railway system: The potential benefits to geographers of harnessing inputs from "naive" outsiders, *Journal of Transport Geography*, 33, pp. 85–94.

Farber, S., L. Fu (2017) Dynamic public transit accessibility using travel time cubes: Comparing the effects of infrastructure (dis)investments over time, *Computers, Environment and Urban Systems*, 62, pp. 30–40.

Farber, S., M. Z. Morang, M. J. Widener (2014) Temporal variability in transit-based accessibility to supermarkets, *Applied Geography*, 53, pp. 149–159.

Farber, S., B. Ritter, L. Fu (2016) Space-time mismatch between transit service and observed travel patterns in the Wasatch Front, Utah: A social equity perspective, *Travel Behaviour and Society*, 4, pp. 40–48.

Farzin, J. (2008) Constructing an Automated Bus Origin-Destination Matrix Using Farecard and Global Positioning System Data in São Paulo, Brazil, *Transportation Research Record: Journal of the Transportation Research Board*, 2072, pp. 30–37.

Fayyaz S., S. K., X. C. Liu, G. Zhang (2017) An efficient General Transit Feed Specification (GTFS) enabled algorithm for dynamic transit accessibility analysis, *PLoS ONE*, 12(10), p. e0185333.

Feng, J., X. Li, B. Mao, Q. Xu, Y. Bai (2017) Weighted complex network analysis of the Beijing subway system: Train and passenger flows, *Physica A: Statistical Mechanics and its Applications*, 474, pp. 213–223.

Freeman, L. C. (1977) A Set of Measures of Centrality Based on Betweenness, *Sociometry*, 40(1), pp. 35–41.

Gallotti, R., M. Barthelemy (2015) The multilayer temporal network of public transport in Great Britain, *Scientific Data*, 2, p. 140056.

Gao, S., Y. Wang, Y. Gao, Y. Liu (2013) Understanding urban traffic-flow characteristics: A rethinking of betweenness centrality, *Environment and Planning B: Planning and Design*, 40(1), pp. 135–153.

Gentile, G., M. Florian, Y. Hamdouch, O. Cats, A. Nuzzolo (2016) The theory of transit assignment: basic modelling frameworks, in: Gentile, G., K. Noekel, eds., *Modelling Public Transport Passenger Flows in the Era of Intelligent Transport Systems*, chap. 6, Springer, pp. 287–386.

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.

Google (2019) GTFS Static Overview, URL `https://developers.google.com/transit/gtfs/`.

Gordon, J., H. Koutsopoulos, N. Wilson, J. Attanucci (2013) Automated Inference of Linked Transit Journeys in London Using Fare-Transaction and Vehicle Location Data, *Transportation Research Record: Journal of the Transportation Research Board*, 2343, pp. 17–24.

Goulet Langlois, G., H. N. Koutsopoulos, J. Zhao, G. G. Langlois, 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.

Guo, Z. (2011) Mind the map! The impact of transit maps on path choice in public transit, *Transportation Research Part A: Policy and Practice*, 45(7), pp. 625–639.

Hagberg, A., D. S. Chult., P. Swart (2008) Exploring network structure, dynamics, and function using NetworkX, in: Varoquaux, G., T. Vaught, J. Millman, eds., *Proceedings of the 7th Python in Science Conference (SciPy2008)*, Pasadena, CA, USA, pp. 11–15.

Hassan, M. N., T. H. Rashidi, S. T. Waller, N. Nassir, M. Hickman (2016) Modeling Transit Users Stop Choice Behavior: Do Travelers Strategize?, *Journal of Public Transportation*, 19(3), pp. 98–116.

Haznagy, A., I. Fi, A. London, T. Nemeth (2015) Complex network analysis of public transportation networks: A comprehensive study, in: *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, IEEE, Budapest, Hungary, pp. 371–378.

Hillier, B., A. Penn, J. Hanson, T. Grajewski, J. Xu (1993) Natural movement: or, configuration and attraction in urban pedestrian movement, *Environment and Planning B: Planning and Design*, 20(1), pp. 29–66.

Hörcher, D., D. J. Graham, R. J. Anderson (2017) Crowding cost estimation with large scale smart card and vehicle location data, *Transportation Research Part B: Methodological*, 95, pp. 105–125.

Hotelling, H. (1933) Analysis of a complex of statistical variables into principal components, *Journal of Educational Psychology*, 24(6), pp. 417–441.

Hsiao, C. (2007) Panel data analysis-advantages and challenges, *Test*, 16(1), pp. 1–22.

Ji, Y., R. G. Mishalani, M. R. McCord (2014) Estimating Transit Route OD Flow Matrices from APC Data on Multiple Bus Trips Using the IPF Method with an Iteratively Improved Base: Method and Empirical Evaluation, *Journal of Transportation Engineering*, 140(5), p. 04014008.

Ji, Y., R. G. Mishalani, M. R. McCord (2015) Transit passenger origin–destination flow estimation: Efficiently combining onboard survey and large automatic passenger

count datasets, *Transportation Research Part C: Emerging Technologies*, 58, pp. 178–192.

Jiang, B., C. Liu (2009) Street-based topological representations and analyses for predicting traffic flow in GIS, *International Journal of Geographical Information Science*, 23(9), pp. 1119–1137.

Kazerani, A., S. Winter (2009) Can betweenness centrality explain traffic flow?, in: *12th AGILE International Conference on Geographic Information Scienc*, Hanover, Germany, pp. 1–9.

Kieu, L. M., A. Bhaskar, E. Chung (2015a) A modified Density-Based Scanning Algorithm with Noise for spatial travel pattern analysis from Smart Card AFC data, *Transportation Research Part C: Emerging Technologies*, 58, pp. 193–207.

Kieu, L. M., A. Bhaskar, E. Chung (2015b) Passenger segmentation using smart card data, *IEEE Transactions on Intelligent Transportation Systems*, 16(3), pp. 1537–1548.

Kivelä, M., A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, M. A. Porter (2014) Multilayer networks, *Journal of Complex Networks*, 2(3), pp. 203–271.

Koutsopoulos, H., P. Noursalehi, Y. Zhu, N. Wilson (2017) Automated data in transit: Recent developments and applications, in: *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, Naples, pp. 604–609.

Koutsopoulos, H. N., Z. Ma, P. Noursalehi, Y. Zhu (2019) Chapter 10 - Transit Data Analytics for Planning, Monitoring, Control, and Information, in: *Mobility Patterns, Big Data and Transport Analytics*, Elsevier, pp. 229–261.

Kujala, R., C. Weckstrom, R. K. Darst, M. N. Mladenovic, J. Saramaki (2018a) Data Descriptor: A collection of public transport network data sets for 25 cities, *Scientific Data*, 5, p. 180089.

Kujala, R., C. Weckström, M. N. Mladenović, J. Saramäki (2018b) Travel times and transfers in public transport: Comprehensive accessibility analysis based on Pareto-optimal journeys, *Computers, Environment and Urban Systems*, 67, pp. 41–54.

Kumar, P., A. Khani, Q. He (2018) A robust estimation of transit passenger trajectories using automated data, *Transportation Research Part C: Emerging Technologies*, 95, pp. 731–747.

Kusakabe, T., T. Iryo, Y. Asakura (2010) Estimation method for railway passengers' train choice behavior with smart card transaction data, *Transportation*, 37(5), pp. 731–749.

Lakhina, A., K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, N. Taft (2004) Structural Analysis of Network Traffic Flows, *SIGMETRICS Perform. Eval. Rev.*, 32(1), pp. 61–72.

Latora, V., M. Marchiori (2001) Efficient Behavior of Small-World Networks, *Physical Review Letters*, 87, p. 198701.

Latora, V., M. Marchiori (2002) Is the Boston subway a small-world network?, *Physica A: Statistical Mechanics and its Applications*, 314(1-4), pp. 109–113.

Lee, S., M. Hickman, D. Tong (2012) Stop Aggregation Model:Development and Application, *Transportation Research Record: Journal of the Transportation Research Board*, 2276, pp. 38–47.

Lee, S., M. Hickman, D. Tong (2013) Development of a temporal and spatial linkage between transit demand and land-use patterns, *Journal of Transport and Land Use*, 6(2), pp. 33–46.

Lee, S. G., M. Hickman (2013) Are Transit Trips Symmetrical in Time and Space?, *Transportation Research Record: Journal of the Transportation Research Board*, 2382, pp. 173–180.

Lei, T. L., R. L. Church (2010) Mapping transit-based access: Integrating GIS, routes and schedules, *International Journal of Geographical Information Science*, 24(2), pp. 283–304.

Lin, J., Y. Ban (2013) Complex Network Topology of Transportation Systems, *Transport Reviews*, 33(6), pp. 658–685.

Liu, Y., J. Bunker, L. Ferreira (2010) Transit users' route-choice modelling in transit assignment: A review, *Transport Reviews*, 30(6), pp. 753–769.

Louf, R., C. Roth, M. Barthelemy (2014) Scaling in transportation networks, *PLoS ONE*, 9(7), p. 102007.

Luo, D., L. Bonnetain, O. Cats, H. van Lint (2018) Constructing Spatiotemporal Load Profiles of Transit Vehicles with Multiple Data Sources, *Transportation Research Record*, 2672(8), pp. 175–186.

Luo, D., O. Cats, H. van Lint (2019) Can passenger flow distribution be estimated solely based on network properties in public transport systems?, *Transportation*, pp. https://doi.org/10.1007/s11116–019–09990–w.

Ma, X., Y.-J. Wu, Y. Wang, F. Chen, J. Liu (2013) Mining Smart Card Data for Transit Raiders' Travel Patterns, *Transportation Research Board, 92nd Annual Meeting*, 36, pp. 1–12.

Ma, Z., H. N. Koutsopoulos, L. Ferreira (2017) Quantile Regression Analysis of Transit Travel Time Reliability Using Automatic Vehicle Location and Fare Card Data, *Transportation Research Record: Journal of the Transportation Research Board*, 2652, pp. 19–29.

Ma, Z.-L., L. Ferreira, M. Mesbah, A. T. Hojati (2015) Modeling Bus Travel Time Reliability with Supply and Demand Data from Automatic Vehicle Location and Smart Card Systems, *Transportation Research Record: Journal of the Transportation Research Board*, 2533, pp. 17–27.

Marquardt, D. W. (1980) Comment: You should standardize the predictor variables in your regression models, *Journal of the American Statistical Association*, 75(369), pp. 87–91.

McCord, M., R. Mishalani, X. Hu (2012) Grouping of Bus Stops for Aggregation of Route-Level Passenger Origin-Destination Flow Matrices, *Transportation Research Record: Journal of the Transportation Research Board*, 2277, pp. 38–48.

McQuenn, J. (1967) Some methods for classification and analysis of multivariate observations, in: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297.

Min, Y. H., S. J. Ko, K. M. Kim, S. P. Hong (2016) Mining missing train logs from Smart Card data, *Transportation Research Part C: Emerging Technologies*, 63, pp. 170–181.

Moreira-Matias, L., O. Cats (2016) Toward a Demand Estimation Model Based on Automated Vehicle Location, *Transportation Research Record: Journal of the Transportation Research Board*, 2544, pp. 141–149.

Moreira-Matias, L., J. Mendes-Moreira, J. F. De Sousa, J. Gama (2015) Improving Mass Transit Operations by Using AVL-Based Systems: A Survey, *IEEE Transactions on Intelligent Transportation Systems*, 16(4), pp. 1636–1653.

Munizaga, M., F. Devillaine, C. Navarrete, D. Silva (2014) Validating travel behavior estimated from smartcard data, *Transportation Research Part C: Emerging Technologies*, 44, pp. 70–79.

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.

Musso, A., V. R. Vuchic (1988) Characteristics of metro networks and methodology for their evaluation, *Transportation Research Record*, 1162, pp. 22–33.

Nassir, N., M. Hickman, A. Malekzadeh, E. Irannezhad (2015) Modeling Transit Passenger Choices of Access Stop, *Transportation Research Record: Journal of the Transportation Research Board*, 2493, pp. 70–77.

Nassir, N., M. Hickman, A. Malekzadeh, E. Irannezhad (2016) A utility-based travel impedance measure for public transit network accessibility, *Transportation Research Part A: Policy and Practice*, 88, pp. 26–39.

Nassir, N., A. Khani, S. Lee, H. Noh, M. Hickman (2011) Transit Stop-Level Origin-Destination Estimation Through Use of Transit Schedule and Automated Data Collection System, *Transportation Research Record: Journal of the Transportation Research Board*, 2263, pp. 140–150.

Newman, M. (2010) *Networks: An introduction*, Oxford University Press.

Nguyen, S., S. Pallottino (1988) Equilibrium traffic assignment for large scale transit networks, *European Journal of Operational Research*, 37(2), pp. 176–186.

Noursalehi, P., H. N. Koutsopoulos, J. Zhao (2018) Real time transit demand prediction capturing station interactions and impact of special events, *Transportation Research Part C: Emerging Technologies*, 97, pp. 277–300.

Nuzzolo, A., F. Russo, U. Crisalli (2001) A Doubly Dynamic Schedule-based Assignment Model for Transit Networks, *Transportation Science*, 35(3), pp. 268–285.

Ortúzar, J. D., L. G. Willumsen (2011) *Modelling Transport*, 4 ed., John Wiley & Sons.

Pearson, K. (1901) LIII. On lines and planes of closest fit to systems of points in space, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), pp. 559–572.

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.

Penn, A., B. Hillier, D. Banister, J. Xu (1998) Configurational modelling of urban movement networks, *Environment and Planning B: Planning and Design*, 25(1), pp. 59–84.

Puzis, R., Y. Altshuler, Y. Elovici, S. Bekhor, Y. Shiftan, A. Pentland (2013) Augmented betweenness centrality for environmentally aware traffic monitoring in transportation networks, *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations*, 17(1), pp. 91–105.

Ray, S., R. Turi (1999) Determination of number of clusters in k-means clustering and application in colour image segmentation, *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pp. 137–143.

Roth, C., S. M. Kang, M. Batty, M. Barthelemy (2012) A long-time limit for world subway networks, *Journal of the Royal Society Interface*, 9(75), pp. 2540–2550.

Saghapour, T., S. Moridpour, R. G. Thompson (2016) Public transport accessibility in metropolitan areas: A new approach incorporating population density, *Journal of Transport Geography*, 54, pp. 273–285.

Schmöcker, J. D., A. Fonzone, H. Shimamoto, F. Kurauchi, M. G. Bell (2011) Frequency-based transit assignment considering seat capacities, *Transportation Research Part B: Methodological*, 45(2), pp. 392–408.

Sen, P., S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, S. S. Manna (2003) Small-world properties of the Indian railway network, *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 67(3), p. 5.

Sienkiewicz, J., J. A. Holyst (2005) Statistical analysis of 22 public transport networks in Poland, *Physical Review E*, 72(4), p. 46127.

Soh, H., S. Lim, T. Zhang, X. Fu, G. K. K. Lee, T. G. G. Hung, P. Di, S. Prakasam, L. Wong (2010) Weighted complex network analysis of travel routes on the Singapore public transportation system, *Physica A: Statistical Mechanics and its Applications*, 389(24), pp. 5852–5863.

Spiess, H., M. Florian (1989) Optimal strategies: a new assignment model for transit networks, *Transportation Research Part B: Methodological*, 23(2), pp. 83–102.

Sun, L., K. W. Axhausen (2016) Understanding urban mobility patterns with a probabilistic tensor factorization framework, *Transportation Research Part B: Methodological*, 91, pp. 511–524.

Sun, L., K. W. Axhausen, D.-H. Lee, X. Huang (2013) Understanding metropolitan patterns of daily encounters, *Proceedings of the National Academy of Sciences*, 110(34), pp. 13774–13779.

Sun, L., D.-H. Lee, A. Erath, X. Huang (2012) Using smart card data to extract passenger's spatio-temporal density and train's trajectory of MRT system, in: *Proceedings of the ACM SIGKDD international workshop on urban computing*, ACM, pp. 142–148.

Tamblay, S., P. Galilea, P. Iglesias, S. Raveau, J. C. Muñoz (2016) A zonal inference model based on observed smart-card transactions for Santiago de Chile, *Transportation Research Part A: Policy and Practice*, 84, pp. 44–54.

Tou, J. T., R. C. Gonzalez (1974) *Pattern recognition principles*, Massachusetts: Addison-Wesley.

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: Technology, Planning, and Operations*, 11(1), pp. 1–14.

Turner, A. (2007) From axial to road-centre lines: A new representation for space syntax and a new model of route choice for transport network analysis, *Environment and Planning B: Planning and Design*, 34(3), pp. 539–555.

Utsunomiya, M., J. Attanucci, N. Wilson (2006) Potential uses of transit smart card registration and transaction data to improve transit planning, *Transportation Research Record: Journal of the Transportation Research Board*, 1971, pp. 119–126.

van Nes, R. (2002) *Design of multimodal transport networks: A hierarchical approach*, Ph.D. thesis, Delft University of Technology.

van Nes, R., R. Hamerslag, B. H. Immers (1988) Design of public transport networks, *Transportation Research Record*, 1202, pp. 74–83.

van Oort, N., T. Brands, E. de Romph (2015a) Short-Term Prediction of Ridership on Public Transport with Smart Card Data, *Transportation Research Record: Journal of the Transportation Research Board*, 2535, pp. 105–111.

van Oort, N., D. Sparing, T. Brands, R. M. Goverde (2015b) Data driven improvements in public transport: the Dutch example, *Public Transport*, 7(3), pp. 369–389.

van Wee, B. (2016) Accessible accessibility research challenges, *Journal of Transport Geography*, 51, pp. 9–16.

Verleysen, M., D. François (2011) The Curse of Dimensionality in Data Mining and Time Series Prediction, *Proceedings of the 8th International Conference on Artificial Neural Networks: Computational Intelligence and Bioinspired Systems*, pp. 758–770.

Vlahogianni, E. I., B. B. Park, H. van Lint (2015) Big data in transportation and traffic engineering, *Transportation Research Part C: Emerging Technologies*, 58, p. 161.

von Ferber, C., T. Holovatch, Y. Holovatch, V. Palchykov (2007) Network harness: Metropolis public transport, *Physica A: Statistical Mechanics and its Applications*, 380(1-2), pp. 585–591.

von Ferber, C., T. Holovatch, Y. Holovatch, V. Palchykov (2009) Public transport networks: empirical analysis and modeling, *The European Physical Journal B*, 68, pp. 261–275.

Vuchic, V., A. Musso (1991) Theory and practice of metro network design, *Public Transport International*, 40(3), p. 298.

Vuchic, V. R. (2005) Transit Lines and Networks, in: *Urban Transit: Operations, Planning, and Economics*, chap. 4, John Wiley & Sons.

Wang, W., J. Attanucci, N. Wilson (2011) Bus Passenger Origin-Destination Estimation and Related Analyses Using Automated Data Collection Systems, *Journal of Public Transportation*, 14(4), pp. 131–150.

Watts, D. J., S. H. Strogatz (1998) Collective dynamics of 'small-world' networks, *Nature*, 393(6684), pp. 440–442.

Welch, T. F., A. Widita (2019) Big data in public transportation: a review of sources and methods, *Transport Reviews*, p. DOI: 10.1080/01441647.2019.1616849.

Wen, T. H., W. C. B. Chin, P. C. Lai (2017) Understanding the topological characteristics and flow complexity of urban traffic congestion, *Physica A: Statistical Mechanics and its Applications*, 473, pp. 166–177.

Wilson, N. H. M., J. Zhao, A. Rahbee (2009) The potential impact of automated data collection systems on urban public transport planning., in: *Schedule-Based Modeling of Transportation Networks: Theory and Applications*, Springer, pp. 75–99.

Wong, J. (2013) Leveraging the General Transit Feed Specification for efficient transit analysis, *Transportation Research Record: Journal of the Transportation Research Board*, 2338, pp. 11–19.

Xu, X., J. Hu, F. Liu, L. Liu (2007) Scaling and correlations in three bus-transport networks of China, *Physica A: Statistical Mechanics and its Applications*, 374(1), pp. 441–448.

Yan, F., C. Yang, S. V. Ukkusuri (2019) Alighting stop determination using two-step algorithms in bus transit systems, *Transportmetrica A: Transport Science*, 15(2), pp. 1522–1542.

Yap, M., O. Cats, B. van Arem (2018) Crowding valuation in urban tram and bus transportation based on smart card data, *Transportmetrica A: Transport Science*, p. DOI: 10.1080/23249935.2018.1537319.

Ye, P., B. Wu, W. Fan (2016) Modified Betweenness-Based Measure for Prediction of Traffic Flow on Urban Roads, *Transportation Research Record: Journal of the Transportation Research Board*, 2563, pp. 144–150.

Zanin, M., X. Sun, S. Wandelt (2018) Studying the Topology of Transportation Systems through Complex Networks : Handle with Care, *Journal of Advanced Transportation*, 2018, p. Article ID 3156137.

Zhang, Y., W. H. K. Lam, A. Sumalee, H. K. Lo, C. O. Tong (2010) The multi-class schedule-based transit assignment model under network uncertainties, *Public Transport*, 2(1), pp. 69–86.

Zhao, J., F. Zhang, L. Tu, C. Xu, D. Shen, C. Tian, X. Y. Li, Z. Li (2017a) Estimation of Passenger Route Choice Pattern Using Smart Card Data for Complex Metro Systems, *IEEE Transactions on Intelligent Transportation Systems*, 18(4), pp. 790–801.

Zhao, S., P. Zhao, Y. Cui (2017b) A network centrality measure framework for analyzing urban traffic flow: A case study of Wuhan, China, *Physica A: Statistical Mechanics and its Applications*, 478, pp. 143–157.

Zhong, C., E. Manley, S. Müller Arisona, M. Batty, G. Schmitt (2015) Measuring variability of mobility patterns from multiday smart-card data, *Journal of Computational Science*, 9, pp. 125–130.

Zhou, Y., L. Yao, Y. Chen, Y. Gong, J. Lai (2017) Bus arrival time calculation model based on smart card data, *Transportation Research Part C: Emerging Technologies*, 74, pp. 81–96.

Zhu, Y., H. N. Koutsopoulos, N. H. Wilson (2017) A probabilistic Passenger-to-Train Assignment Model based on automated data, *Transportation Research Part B: Methodological*, 104, pp. 522–542.

Summary

Public transport (PT) plays an increasingly important role in solving mobility challenges, especially in densely populated metropolitan areas. Further improving PT systems requires more advanced planning and operations. Fortunately, the considerable amount of data that have become increasingly available for PT systems offer an opportunity to address this challenge. However, how these data can be effectively used to achieve this goal still remains as an unresolved question in the scientific literature. More research is therefore needed to bridge this gap in order to advance PT systems for addressing mobility challenges.

To this end, this dissertation is focused on developing methods and models for translating high-volume data from various sources into novel knowledge and insights that can be used to improve PT planning and operations. These data sources consist of automatic fare collection (AFC), automatic vehicle location (AVL), and general transit feed specification (GTFS) data. We propose a framework for processing and fusing these data, which allows for obtaining information key to PT research and applications. Moreover, we present a framework for conducting data-driven PT research, the ultimate goal of which is to contribute to the improvement of planning and operations.

This dissertation first examines how to obtain onboard occupancy of PT vehicles by integrating all the three different data sources mentioned above. We first specifically identify the issues related to each and the combination of different data sources. Then based on this diagnosis, we propose a method for systematically addressing these issues, which results in desired profiles of PT vehicles with onboard occupancy and improved trajectories. We demonstrate the proposed method using the data from the PT system of The Hague, the Netherlands. The load profiles are further visualized using space-time seat occupancy diagrams, which provides operators with a compact and powerful reference to intuitively examine the onboard crowding patterns over time and space. This visualization technique can help operators in timetable optimization, network and fleet scheduling, and sub-route service design.

Second, this dissertation deals with the issue of high-dimensionality in large-scale passenger flows. The high-dimensionality makes it challenging to unravel and further model passenger flow dynamics to a large spatiotemporal extent. We hence attempt to address this challenge by applying principal component analysis (PCA), a popular dimensionality reduction technique. We first show how the matrix of multivariate

time series of passenger flows can be constructed, and then specify how such high-dimensional flow matrix can be transformed using PCA. To demonstrate the method, we employ the metro system of Shenzhen, China for the case study. The results show that a great amount of variance contained in the original data can be effectively retained in lower-dimensional sub-spaces composed of a few top principal components. We further analyze the features of such low dimensionality, with detailed investigations into the principal components and temporal stability of the flow structure. This study contributes to the understanding of large-scale spatiotemporal PT flow patterns. It also paves the way for integrating PCA into large-scale passenger flow modeling and predictions.

Third, we propose a k-means-based method to cluster PT stops for constructing zone-to-zone OD matrices. The key is to heuristically determining the number of clusters by combining spatial distances and passenger OD flows. Differing from the traditional way of grouping stops based on predefined traffic analysis zones, our proposed method provides a data-driven perspective for solving such problems using passenger flows that can be directly observed rather than their proxies. The method is demonstrated by a case study of the PT system of The Hague, the Netherlands. We adopt the criterion that the ratio of average intra-cluster flow to average inter-cluster flow should be maximized while maintaining the spatial compactness of all clusters. This clustering approach is particularly suitable for urban areas with high density of PT stops, because it allows for integrating travelers' (origin and destination) stop and route choices in subsequent modeling studies.

Fourth, this dissertation presents a new method for analyzing the accessibility of PT service networks based on a novel network science approach. It allows for fast comparative assessment across PT networks. Measuring the accessibility based on the average travel impedance, we propose an innovative weighted graph representation of public transport networks (PTNs) that explicitly incorporates the travel costs determined by the scheduled service attributes contained in GTFS data. Consequently, the method enables efficient computation of minimal generalized travel costs between stop pairs. Such cost is comprised of initial and transfer waiting times, in-vehicle travel times and time-equivalent transfer penalty costs. To demonstrate the high transferability and efficiency of the proposed method, we present a case study assessing worldwide tram networks' accessibility. The results provide new insights into PTN design, benchmark and planning.

Last, we investigate whether passenger flow distribution can be estimated solely based on network properties in PT systems. We study in a reverse-engineering fashion, directly examining the relation between observed flow distributions and network properties. Concepts and methods from network science, including the topological representation of PT infrastructure and service networks and centrality indicators, are applied to provide a systematic and concise way for quantifying PT network properties. All the employed centrality indicators are also interpreted in the context of PT systems, which enriches the application of network science in the transport research. We con-

clude that the selected network properties can indeed be used to approximate the global passenger flow distribution across the network to a reasonable extent of accuracy using solely regression models. However, it should be noted that this conclusion does not imply causality. This is because it is likely that supply provision has been designed to correspond to demand patterns and therefore reflects the interplay between demand and supply distributions. This finding can lead to the development of (i) parsimonious data-driven PT assignment models and (ii) global metrics for overseeing large-scale flow dynamics and the gap between flows and service supplies.

Overall, this dissertation makes multiple scientific contributions to data-driven research on PT systems concerning passenger flows and service networks. With rapidly increasing desire of evidence-based decision making, this dissertation can therefore serve as a timely guidance for practitioners to develop more advanced PT systems capitalizing on the current data richness.

Samenvatting

Het openbaar vervoer (ov) speelt een steeds belangrijkere rol bij het oplossen van de mobiliteitsproblemen, met name in dichtbevolkte stedelijke gebieden. Voor verdere verbetering van de ov-systemen zijn geavanceerdere planning en bedrijfsvoering nodig. Gelukkig komen er echter steeds meer data beschikbaar uit ov-systemen, en die bieden een kans om hier verder in te komen. De wetenschappelijke literatuur heeft echter nog geen antwoord op de vraag hoe deze data effectief kunnen worden ingezet. Er is dus meer onderzoek nodig om deze kloof te overbruggen, en ov-systemen zodanig te verbeteren dat mobiliteitsproblemen kunnen worden aangepakt.

Dit proefschrift is dan ook gericht op het ontwikkelen van methoden en modellen voor het vertalen van grote hoeveelheden data uit verschillende bronnen naar nieuwe kennis en inzichten die gebruikt kunnen worden om planning en bedrijfsvoering van het ov te verbeteren. Het betreft gegevens uit geautomatiseerde kaartverkoop (automatic fare collection, AFC), automatische plaatsbepaling van voertuigen (automatic vehicle location, AVL) en general transit feed specification (GTFS). We stellen een kader voor waarmee deze gegevens kunnen worden verwerkt en samengevoegd, zodat informatie verkregen kan worden die belangrijk is voor ov-onderzoek en -toepassingen. Bovendien presenteren we een kader voor het uitvoeren van datagestuurd ov-onderzoek, met als uiteindelijk doel een bijdrage te leveren aan de verbetering van planning en bedrijfsvoering.

In dit proefschrift wordt eerst onderzocht hoe gegevens over de bezettingsgraad van ov-voertuigen kunnen worden verkregen door de drie bovengenoemde gegevensbronnen met elkaar te integreren. Eerst geven we aan welke problemen er zijn met elk van de afzonderlijke gegevensbronnen en met de verschillende combinaties. Op basis van deze diagnose stellen we vervolgens een methode voor waarmee deze problemen systematisch worden aangepakt, en die resulteert in de gewenste profielen van ov-voertuigen met bezettingsgraad en verbeterde trajecten. We demonstreren de voorgestelde methode aan de hand van gegevens uit het ov-systeem van Den Haag. De belastingsprofielen worden verder gevisualiseerd aan de hand van ruimte-tijddiagrammen van de zitplaatsbezetting. Dit geeft een compacte en krachtige referentie om op intuïtieve wijze de druktepatronen te onderzoeken in de ruimte en de tijd. Met behulp van deze visualisatietechniek kunnen dienstregelingen worden geoptimaliseerd, netwerken en rijdend materieel worden gepland en subroutediensten worden ontworpen.

Ten tweede gaat dit proefschrift in op de problematiek van de hoge dimensionaliteit

van grootschalige passagiersstromen. Deze hoge dimensionaliteit maakt het moeilijk om de dynamiek van de passagiersstroom te ontwarren en verder te modelleren voor grote ruimte- en tijdschalen. Dit probleem proberen we aan te pakken door middel van principale-componentenanalyse (PCA), een populaire techniek om dimensionaliteit te verminderen. We laten eerst zien hoe de matrix van multivariate tijdreeksen van passagiersstromen kan worden geconstrueerd, en vervolgens specificeren we hoe een dergelijke hoogdimensionale stroommatrix kan worden getransformeerd met behulp van PCA. Om deze methode te demonstreren gebruiken we een casestudy van het metronetwerk van de Chinese stad Shenzhen. De resultaten tonen aan dat veel variantie in de oorspronkelijke gegevens effectief kan worden behouden in subruimten met lagere dimensies, die zijn samengesteld uit enkele belangrijke principale componenten. We analyseren verder de kenmerken van een dergelijke lage dimensionaliteit, met gedetailleerd onderzoek naar de principale componenten en de temporele stabiliteit van de stromingsstructuur. Dit onderzoek draagt bij tot een beter begrip van grootschalige ruimtelijk-temporele ov-stroompatronen. Ook legt het een basis voor de integratie van PCA in het modelleren en voorspellen van grootschalige passagiersstromen.

Ten derde stellen we een op *k*-means gebaseerde methode voor om ov-haltes te clusteren voor de constructie van zone-tot-zone herkomst-bestemmingsmatrices. Centraal hierin staat het heuristisch bepalen van het aantal clusters door ruimtelijke afstanden en HB-passagiersstromen te combineren. Terwijl traditioneel haltes worden gegroepeerd op basis van vooraf gedefinieerde verkeersanalysezones, biedt onze voorgestelde methode een datagestuurd perspectief om dergelijke problemen op te lossen met behulp van direct waarneembare passagiersstromen in plaats van hun proxy's. We demonstreren deze methode met een casestudy van het ov-systeem van Den Haag. We hanteren het criterium dat de verhouding tussen de gemiddelde intraclusterstroom en de gemiddelde interclusterstroom moet worden gemaximaliseerd, met behoud van de ruimtelijke compactheid van alle clusters. Deze clustervorming is vooral geschikt voor stedelijke gebieden met een hoge dichtheid aan ov-haltes, omdat het mogelijk is om keuzes van reizigers voor halte en route (voor herkomst en bestemming) te integreren in latere modelleringsonderzoeken.

Ten vierde presenteert dit proefschrift een nieuwe methode om de toegankelijkheid van ov-dienstennetwerken te analyseren op basis van een nieuwe aanpak uit de netwerkwetenschap. Hiermee kunnen snel vergelijkende beoordelingen voor meerdere ov-netwerken worden uitgevoerd. We meten de toegankelijkheid op basis van de gemiddelde reisweerstand en stellen op basis daarvan een innovatieve representatie met gewogen grafen van ov-netwerken voor, waarin expliciet de reiskosten zijn opgenomen die worden bepaald door de tariefkenmerken uit de GTFS-data. De methode maakt hiermee een efficiënte berekening van minimale gegeneraliseerde reiskosten tussen halteparen mogelijk. Deze kosten bestaan uit initiële en overstapwachttijden, reistijden in het voertuig en kosten voor overstappenalty's, uitgedrukt in tijd. Om de grote algemene toepasbaarheid en de efficiëntie van de voorgestelde methode aan te tonen, presenteren we een casestudy waarin we wereldwijd de toegankelijkheid van tramnetwerken beo-

ordelen. De resultaten geven nieuwe inzichten in het ontwerp, de benchmarking en de planning van ov-netwerken.

Tot slot onderzoeken we of de verdeling van passagiersstromen kan worden geschat op basis van de netwerkeigenschappen van ov-systemen. Dit doen op een reverse-engineering-achtige manier, waarbij we direct de relatie tussen de waargenomen stroomverdelingen en de netwerkeigenschappen onderzoeken. We passen concepten en methoden uit de netwerkwetenschap toe, zoals de topologische representatie van ov-infrastructuur en dienstennetwerken en centraliteitsindicatoren, met als resultaat een systematische en beknopte manier om ov-netwerkeigenschappen te kwantificeren. Alle gebruikte centraliteitsindicatoren worden ook geïnterpreteerd in de context van ov-systemen, wat een verrijking betekent voor de toepassing van netwerkwetenschap in transportonderzoek. We concluderen dat de geselecteerde netwerkeigenschappen inderdaad kunnen worden gebruikt om, uitsluitend met behulp van regressiemodellen en met een redelijke mate van nauwkeurigheid, de globale verdeling van passagiersstromen binnen het netwerk te benaderen. Er dient echter te worden opgemerkt dat deze conclusie geen causaliteit impliceert. De reden hiervoor is dat het aanbod waarschijnlijk zo is ontworpen dat het overeenkomt met de vraagpatronen en dus de wisselwerking tussen vraag- en aanbodverdelingen weerspiegelt. Deze bevinding kan leiden tot de ontwikkeling van (i) sobere datagestuurde ov-toewijzingsmodellen en (ii) globale metrieken die inzicht geven in grootschalige stromingsdynamiek en de kloof tussen stromen en dienstverlening.

In algemene zin draagt dit proefschrift in meerdere opzichten bij aan datagestuurd onderzoek naar ov-systemen met betrekking tot passagiersstromen en dienstennetwerken. Bij de snel toenemende behoefte aan evidencebased besluitvorming kan dit proefschrift daarom van pas komen als leidraad voor mensen die werkzaam zijn in de sector, en die geavanceerdere ov-systemen willen ontwikkelen met behulp van de huidige rijkdom aan data.

概述

公共交通对解决人们的出行问题起着至关重要的作用，尤其是对人口密集的都市区域。为了打造更加先进智能的公交系统来克服这一挑战，我们需要提升公共交通的规划和运营水平。而日益增长的数据资源为此提供了绝佳的机会。但是，如何有效地利用这些数据资源来推进公共交通规划和运营的升级仍然是目前学术界中尚未完全解决的问题。该博士研究课题因此致力于弥补这一学术空缺。

该博士研究聚焦于开发基于数据驱动的客流和服务网络的分析算法以及数学模型。这些算法和模型可以将海量多源数据转化为对复杂公交系统更为深刻的理解，以此来帮助实现更加智能的公共交通规划以及运营。我们使用的数据包括智能卡（AFC），车辆定位（AVL）以及通用公交咨询规格（GTFS）数据。基于这些数据，我们首先提出了处理和融合多源数据的框架，以及利用这些信息进行以提升规划和运营为目的的数据驱动的研究框架。围绕提出的框架，整个博士论文从涉及出行需求以及出行供给的两个角度来开展，具体的研究包括以下五个方面：

利用三种数据源推断公共交通车辆的拥挤程度

不同数据源之间无法做到完美匹配，需要算法来进行匹配，以实现最佳数据融合。作为之后研究的基础，该项工作解决了一个十分实际的问题。我们设计出一套流程，确保不同数据在经过流程之后可以得出更简约重要的信息，包含了每辆公交车上的载客量，即拥挤程度。我们还提出一种创新的可视化方法，可以直观地识别出过度拥挤的区域以及时间。

借助主成分分析法理解客流在大范围时空中的动态变化

公共交通系统具有极强的时空特性。在需求方面，乘客可以不同时间从成百上千个站点进入公共交通系统，如何有效分析客流特性变得十分棘手。我们通过主成分分析法来降低客流数据的维度，发现将客流的变化投射到另外一个线性空间之后可以有效减少所需分析的变量的数量。这为之后进行短期预测及大规模客流建模提供了一些启示。

提出一种空间聚类方法来合并公共交通站点以建立更集聚的出行起讫矩阵

为了建立更加集聚的公共交通起讫出行矩阵，将公共交通站点进行聚类是较为常见的做法。然而，传统的方法局限于定性地分割网络。为了将额外的客流信息融入进来，我们提出一种数据驱动的算法。我们利用k-means算法进行空间分

割，并利用客流信息来辅助决定最佳的簇数量。这种方法有效利用了可以观测到的信息，为学界提供了一种新的解决问题的思路。

提出一种基于复杂网络理论的评估公共交通可达性的方法

我们提出了一种基于复杂网络理论来评估公共交通可达性的方法。我们将可达性通过旅行阻抗指标来量化，并将其定义为广义旅行成本，包括乘车时间，换乘时间，以及换乘惩罚。为了实现高效计算，我们提出了一种新型的公共交通网络的权重拓扑表达，权重全部来自GTFS数据。我们将该方法应用于评估8个来自世界各地的电车网络。进一步分析表明不同的出行组成部分解释了可达性的变化。这种评估方法对提升公共交通服务网络规划水平具有很大的帮助，允许我们从大量的网络中吸取经验，进行对比。

首次定性研究了公共交通客流分布与网络属性的关系

这是一项探索开创性的研究。我们试图理解公共交通系统中客流分布和网络属性的关系。我们所提出的方法利用了逆向工程的思路，直接利用观测到的客流分布。我们利用一系列的网络中心性指标，结合公共交通网络的拓扑关系表达，很好地量化了公共交通网络的基础设施层以及服务层的属性。此外，这些来自于复杂网络理论的指标也很好地在公共交通的背景下得到诠释。我们使用回归模型去捕捉两种变量之间的关系。通过来自荷兰阿姆斯特丹和海牙的公交网络的数据，我们发现我们的假设可以成立：即公共交通系统中客流分布可以被网络属性所估计。但需要指出这并不代表因果关系。

该博士论文展示了如何利用目前丰富的数据资源去更好的理解复杂的公共交通系统，从而实现更加先进的管理、规划和运营。我们的研究实现了多重的方法上的创新，为学术界提供了贡献；同时，我们的研究也为公共交通领域的从业者提供了一份如何充分利用宝贵数据资源的指导方针。

About the Author

Ding Luo (罗丁 in Chinese) was born on February 7th, 1991 in Tianshui, Gansu Province, China. He grew up in his hometown and moved to Beijing for his bachelor's study in August 2009, where he majored in traffic and transport engineering at Beijing Jiaotong University. In the summer of 2011, Ding attended a summer school at Columbia University in New York City, which marked the beginning of his study tour around the world.

In August 2012, Ding completed his bachelor's study in Beijing and continued his master's program in transport and geoinformation technology at KTH Royal Institute of Technology in Stockholm, Sweden. In October 2013, he started working as a research assistant at the department of transport science, focusing on data analytics and modeling of cyclist behavior. In June 2014, he obtained his master's degree with a thesis entitled "Modeling of cyclists acceleration behavior using naturalistic data". Later, he worked as a research assistant at the same department for several more months.

In March 2015, Ding moved to Shenzhen, China and began his first job at Shenzhen Urban Transport Planning Center. He worked on a couple of projects related to transport modeling and analytics during that period. In March 2016, Ding came back to Europe and started his PhD at TU Delft, funded by an EU Horizon 2020 project, SETA. His task was to develop analytical models for public transport systems. During his PhD, Ding conducted a two-month research visit at Monash University in Melbourne, Australia in 2018.

Ding's research interests center around data analytics on mobility systems based on statistical, machine learning and network science methods. After his PhD, Ding is planning to pursue an industry career in the field of data science.

Refereed journal papers

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

[2] Luo, D., Cats, O. & van Lint, H. (2019) Can passenger flow distribution be estimated solely based on network properties in public transport systems? *Transportation*, https://doi.org/10.1007/s11116-019-09990-w

[3] Yap, M., Luo, D., Cats, O., van Oort, N. & Hoogendoorn, S. (2019) Where shall we sync? Clustering passenger flows to identify urban public transport hubs and their key synchronization priorities. *Transportation Research Part C: Emerging Technologies*, 98, 433-448.

[4] Luo, D., Bonnetain, L., Cats, O. & van Lint, H. (2018) Constructing spatiotemporal load profiles of transit vehicles with multiple data sources. *Transportation Research Record*, 2672(8), 175-186.

[5] Luo, D., Cats, O. & van Lint, H. (2017) Constructing transit origin-destination matrices with spatial clustering. *Transportation Research Record*, 2652(1), 39-49.

[6] Ma, X., & Luo, D. (2016) Modeling cyclist acceleration process for bicycle traffic simulation using naturalistic data. *Transportation Research Part F: Traffic Psychology and Behaviour*, 40, 130-144.

Refereed conference full-papers

[1] Yap, M., Luo, D., & Cats, O. (2018) Where shall we sync? Using passenger flows to determine key interchange connections. *Conference on Advanced Systems in Public Transport and TransitData 2018 (CASPT)*, Brisbane, Australia.

[2] Degeler, V., Heydenrijk-Ottens, L., Luo, D., van Oort, N., & van Lint, H. (2018) Unsupervised approach to bunching swings phenomenon analysis. *Conference on Advanced Systems in Public Transport and TransitData 2018 (CASPT)*, Brisbane, Australia.

[3] Luo, D., Bonnetain, L., Cats, O. & van Lint, H. (2018) Constructing spatiotemporal load profiles of transit vehicles with multiple data sources. *Transportation Research Board 97th Annual Meeting (TRB)*, Washington, D.C., USA.

[4] Luo, D., Cats, O. & van Lint, H. (2017) Analysis of network-wide transit passenger flows based on principal component analysis. *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, Naples, Italy.

[5] Luo, D., Cats, O. & van Lint, H. (2017) Constructing transit origin-destination matrices using spatial clustering. *Transportation Research Board 96th Annual Meeting (TRB)*, Washington, D.C., USA.

[6] Luo, D., & Ma, X. (2016) Modeling of cyclist acceleration behavior using naturalistic GPS data. *Transportation Research Board 95th Annual Meeting (TRB)*, Washington, D.C., USA.

[7] Luo, D., & Ma, X. (2014) Analysis of cyclist behavior using naturalistic data: data processing for model development. *3rd International Cycling Safety Conference (ICSC)*, Gothenburg, Sweden.

Conference Presentations

[1] "Data-driven analysis and modeling of passenger flows and service networks for public transport systems". **Oral presentation**. *Transportation Research Board 99th Annual Meeting (TRB)*, Washington, D.C., USA. January, 2020.

[2] "Integrating network science and public transport accessibility analysis for comparative assessment". **Oral presentation**. *5th International Workshop and Symposium on Transit Data*, Paris, France. July, 2019.

[3] "Can passenger flow distribution be estimated based on network properties in public transport systems?". **Oral presentation**. *Conference on Advanced Systems in Public Transport and TransitData 2018 (CASPT)*, Brisbane, Australia. July, 2018.

[4] "Constructing spatiotemporal load profiles of transit vehicles with multiple data sources". **Poster presentation**. *Transportation Research Board 97th Annual Meeting (TRB)*, Washington, D.C., USA. January, 2018.

[5] "Analysis of network-wide transit passenger flows based on principal component analysis". **Oral presentation**. *5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, Naples, Italy. June, 2017.

[6] "Constructing transit origin-destination matrices using spatial clustering". **Oral presentation**. *Transportation Research Board 96th Annual Meeting (TRB)*, Washington, D.C., USA. January, 2017.

[7] "Modeling of cyclist acceleration behavior using naturalistic GPS data". **Poster presentation**. *Transportation Research Board 95th Annual Meeting (TRB)*, Washington, D.C., USA. January, 2016.

TRAIL Thesis Series

The following list contains the most recent dissertations in the TRAIL Thesis Series. For a complete overview of more than 250 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.

Luo, D., Data-driven Analysis and Modeling of Passenger Flows and Service Networks for Public Transport Systems, T2020/2, February 2020, TRAIL Thesis Series, the Netherlands

van Erp, P.B.C., Relative Flow Data: New opportunities for traffic state estimation, T2020/1, February 2020, TRAIL Thesis Series, the Netherlands

Zhu, Y., Passenger-Oriented Timetable Rescheduling in Railway Disruption Management, T2019/16, December 2019, TRAIL Thesis Series, the Netherlands

Chen, L., Cooperative Multi-Vessel Systems for Waterborne Transport, T2019/15, November 2019, TRAIL Thesis Series, the Netherlands

Kerkman, K.E., Spatial Dependence in Travel Demand Models: Causes, implications, and solutions, T2019/14, October 2019, TRAIL Thesis Series, the Netherlands

Liang, X., Planning and Operation of Automated Taxi Systems, T2019/13, September 2019, TRAIL Thesis Series, the Netherlands

Ton, D., Unravelling Mode and Route Choice Behaviour of Active Mode Users, T2019/12, September 2019, TRAIL Thesis Series, the Netherlands

Shu, Y., Vessel Route Choice Model and Operational Model Based on Optimal Control, T2019/11, September 2019, TRAIL Thesis Series, the Netherlands

Luan, X., Traffic Management Optimization of Railway Networks, T2019/10, July 2019, TRAIL Thesis Series, the Netherlands

Hu, Q., Container Transport inside the Port Area and to the Hinterland, T2019/9, July 2019, TRAIL Thesis Series, the Netherlands

Andani, I.G.A., Toll Roads in Indonesia: transport system, accessibility, spatial and equity impacts, T2019/8, June 2019, TRAIL Thesis Series, the Netherlands

Ma, W., Sustainability of Deep Sea Mining Transport Plans, T2019/7, June 2019, TRAIL Thesis Series, the Netherlands

Alemi, A., Railway Wheel Defect Identification, T2019/6, January 2019, TRAIL Thesis Series, the Netherlands

Liao, F., Consumers, Business Models and Electric Vehicles, T2019/5, May 2019, TRAIL Thesis Series, the Netherlands Tamminga, G., A Novel Design of the Transport Infrastructure for Traffic Simulation Models, T2019/4, March 2019, TRAIL Thesis Series, the Netherlands

Lin, X., Controlled Perishable Goods Logistics: Real-time coordination for fresher products, T2019/3, January 2019, TRAIL Thesis Series, the Netherlands

Dafnomilis, I., Green Bulk Terminals: A strategic level approach to solid biomass terminal design, T2019/2, January 2019, TRAIL Thesis Series, the Netherlands

Feng, F., Information Integration and Intelligent Control of Port Logistics System, T2019/1, January 2019, TRAIL Thesis Series, the Netherlands

Beinum, A.S. van, Turbulence in Traffic at Motorway Ramps and its Impact on Traffic Operations and Safety, T2018/12, December 2018, TRAIL Thesis Series, the Netherlands

Bellsolà Olba, X., Assessment of Capacity and Risk: A Framework for Vessel Traffic in Ports, T2018/11, December 2018, TRAIL Thesis Series, the Netherlands

Knapper, A.S., The Effects of using Mobile Phones and Navigation Systems during Driving, T2018/10, December 2018, TRAIL Thesis Series, the Netherlands

Varotto, S.F., Driver Behaviour during Control Transitions between Adaptive Cruise Control and Manual Driving: empirics and models, T2018/9, December 2018, TRAIL Thesis Series, the Netherlands

Stelling-Kończak, A., Cycling Safe and Sound, T2018/8, November 2018, TRAIL Thesis Series, the Netherlands

Essen, van M.A., The Potential of Social Routing Advice, T2018/7, October 2018, TRAIL Thesis Series, the Netherlands

Su, Zhou, Maintenance Optimization for Railway Infrastructure Networks, T2018/6, September 2018, TRAIL Thesis Series, the Netherlands

Cai, J., Residual Ultimate Strength of Seamless Metallic Pipelines with Structural Damage, T2018/5, September 2018, TRAIL Thesis Series, the Netherlands

Ghaemi, N., Short-turning Trains during Full Blockages in Railway Disruption Management, T2018/4, July 2018, TRAIL Thesis Series, the Netherlands

van der Gun, J.P.T., Multimodal Transportation Simulation for Emergencies using the Link Transmission Model, T2018/3, May 2018, TRAIL Thesis Series, the Netherlands