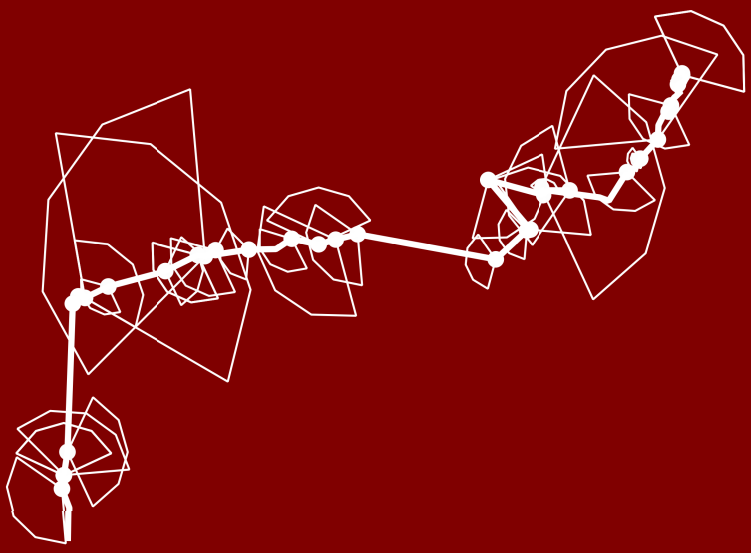
Linköping Studies in Science and Technology
Dissertation No. 2141

Methods for Travel Pattern Analysis Using Large-Scale Passive Data

Nils Breyer



LINKÖPING UNIVERSITY

Linköping Studies in Science and Technology. Thesis №2141
Dissertation

Methods for Travel Pattern Analysis Using Large-Scale Passive Data

Nils Breyer

Department of Science and Technology
Linköping University, SE-601 74 Norrköping, Sweden

Norrköping 2021

Methods for Travel Pattern Analysis Using Large-Scale Passive Data

Nils Breyer

ISBN 978-91-7929-665-0
ISSN 0345-7524

Linköping University
Department of Science and Technology
SE-601 74 Norrköping

Printed by LiU Tryck, Linköping, Sweden 2021

Abstract

Comprehensive knowledge of travel patterns is crucial to enable planning for a more efficient traffic system that accommodates human mobility demand. Currently, this knowledge is mainly based on traffic models based on relatively small samples of observations collected from travel surveys and traffic counts. The data is expensive to collect and provides only partial observations of travel patterns. With the rise of new technology, new large-scale passive data sources can be used to analyse travel patterns. This thesis aims to expand the knowledge about how to use cellular network data collected by cellular network operators and smart-card data from public transit systems to analyse travel patterns. The focus is particularly on the data processing methods needed to extract travel patterns. The thesis's contributions include new methods for extracting trips, estimating travel demand, route inference and travel mode choice from cellular network data and a method to extract travel behaviour changes from smart-card data. Different approaches are proposed to evaluate the methods: the validation using experimental data, validation using other available data sources, and comparison of results obtained using different methods.

The findings include that methods for extracting travel patterns from large-scale passive data need to account for the data's characteristics. Paper II illustrates that route inference from Call Detail Records by strictly following the used cell towers' locations is problematic due to the noise and low resolution of the data. Both rule-based and machine learning methods can be used to extract travel patterns. Paper I shows that a rule-based stop detection algorithm can be used to extract longer trips from cellular network data reliably. On the other hand, Paper III shows that for travel mode classification of trips extracted from cellular network data, supervised classification can outperform rule-based methods. Unsupervised machine learning can be used to find patterns without prior specification. Paper V shows how clustering of smart-card data could be used to group public transit users by travel behaviour to understand the effects of a disruption. Supervised machine learning requires training data. When no or little training data is available, using semi-supervised learning is a promising approach as demonstrated in Paper IV.

In the studies of this thesis, real-world, large-scale passive datasets have been used to demonstrate how the extraction of travel patterns works under realistic circumstances. This has exposed limitations due to the data source's characteristics and limitations due to possible sample bias. At the same time, the studies of this thesis show the potential of using large-scale passive data. Changes in travel patterns can be identified quickly as new data can be collected continuously. Due to the large sample size, the data allows understanding travel patterns based on observations instead of relying on traffic models' underlying assumptions.

Populärvetenskaplig Sammanfattning

Ett effektivt trafiksystem är avgörande för att uppnå klimatmålen och samtidigt tillgodose människors efterfrågan på mobilitet. För att trafikplanerare ska kunna ta välgrundade beslut för att utveckla trafiksystemet krävs en omfattande förståelse av historiska och nuvarande resmönster. Dessa kan sedan användas för att till exempel identifiera persontransporter som kan flyttas till energieffektivare trafikslag eller för att modellera effekterna av en infrastrukturinvestering. Trafikplanerare använder idag trafikmodeller med resvanundersökningar och trafikmätningar som indata. Eftersom dessa datakällor är dyra och innehåller ett mycket begränsat antal observationer kan modellerna endast ge ungefärliga skattningar av resmönster. Nya storskaliga passiva datakällor som data från mobilnätet och data från reskort i kollektivtrafiken öppnar för nya möjligheter att observera resmönster på ett sätt som kan ge en mycket mer detaljerad förståelse av de faktiska resmönstren.

Syftet med den här avhandlingen är att vidga förståelsen för vad som behövs för att processa storskaliga passiva datakällor såsom mobilnätsdata och data från reskort i kollektivtrafiken för att analysera resmönster. Artiklarna i denna avhandling föreslår nya metoder för att detektera resor, estimera reseefterfrågan, skatta ruttval och färdmedelsval från mobilnätsdata samt en metod för att analysera förändringar i resebeteende med data från reskort i kollektivtrafiken. För att bedöma kvalitén på de extraherade resmönstren föreslås olika utvärderingsmetoder: validering med hjälp av experimentella data, validering mot andra datakällor och jämförelse mellan olika metoder. Genom att utvärdera metoderna fås kunskap om potentialen och begränsningarna med att använda storskaliga passiva datakällor för att analysera resmönster.

Studierna i denna avhandling visar på att storskaliga passiva datakällor kan användas för att förstå resmönster på ett mycket mer detaljerat sätt än vad som är möjligt med hjälp av resvanundersökningar och trafikmätningar. Resultaten visar bland annat att mobilnätsdata kan användas för att estimera reseefterfrågan, men att det finns en risk att särskilt korta resor inte detekteras tillförlitligt. Om lågupplöst mobilnätsdata användas för att skatta flöden på vägnätet spelar ruttestimeringsmetoden stor roll. Resultaten när det gäller färdmedelsklassificering av resor från mobilnätet visar att metoder som jämför observationer med tillgängliga ruttalternativ funkar sämre om ruttalternativen för olika färdmedel ligger rumsligt nära. Bättre resultat kan uppnås med maskininlärningsmetoder och det är även möjligt att uppnå bra resultat om ingen träningsdata finns tillgänglig. Den sista studien visar hur data från reskort i kollektivtrafiken kan användas för att analysera förändringar i resmönster efter en långvarig störning i kollektivtrafiksystemet.

Acknowledgments

First, I would like to thank my main supervisor Clas Rydergren and my co-supervisor David Gundlegård. I very much appreciate that you have always been available and gave me guidance and inspiration. Thank you for sharing your knowledge in lengthy discussions and taking the time to give feedback that challenged me to continuously improve. I very much enjoyed working together with you! I would also like to thank Lars Sköld, Simon Moritz, Ida Kristofferson and Di Yuan, and all others who made this work possible with their engagement.

The research in this thesis has been to a large extend funded by the projects Mobile Network Origin Destination Estimation (MODE) and Mobile Network Data in Future Transport Systems (MOFT), both financed by Vinnova and Demand model estimation based on combination of active and passive data collection (DEMOPAN) funded by Trafikverket.

I also want to thank all my colleagues at the division of Communication and Transportation Systems (KTS). I really appreciated being part of a group this international, with different research fields and perspectives. The last year of working from home made it clear how much I miss the interaction and informal discussion with you! In particular, I want to thank the group of PhD students for the great company, fun times and all mutual support. Special thanks go to Niki and Mats for their commitment to represent the PhD students at KTS!

A big thank also to Morten Eltved, Jesper Bláfoss Ingvardson and Otto Anker Nielsen at DTU in Copenhagen. I really enjoyed working with you, even though my research visit turned out to be two and a half only weeks instead of three months, due to the pandemic. Yet, two fredagsøl were enough to start appreciating the Danish culture.

Finally, I would like to thank all my friends, my sister Annalena and my dear parents Marita and Gerd-Herbert, for their unconditional support. Thank you, Fanny, for all your love and for reminding me that there are other things in life than research.

Norrköping, May 2021

Nils Breyer

Contents

Chapter 1

Introduction

An efficient and sustainable traffic system is essential to achieve global climate goals while accommodating human mobility demand. Only with comprehensive knowledge about travel patterns, traffic planners can make informed decisions when developing the traffic system. Travel patterns describe human mobility, including when, why and how people move between different places. With a good understanding of travel patterns, we can estimate the travel demand, that is, the number of people travelling between different areas. This also allows us, for example, to identify travel demand that could be using a more sustainable travel mode. We can also use travel patterns to describe the decision making process of travellers how to travel, also called travel behaviour. Further, a good understanding of travel patterns allows modelling and forecasting the effects of changes to the traffic system. Before making an infrastructure investment such as constructing a new railway or road, we can then analyse its anticipated effects on the traffic system.

In the past, the main sources of observations of travel patterns have been travel surveys and traffic counts. A significant limitation of these data sources is their sample size, limited by the expensive data collection needed. Today, new large-scale passive data sources such as cellular network data and smart card data are available, which open up the possibility to obtain large samples of observations of travel patterns. However, extensive data processing is needed to obtain relevant travel patterns from the raw data. The focus of this thesis is the design of these data processing methods and their evaluation.

1.1 Motivation

The traffic system is a highly complex system with many combinations of origins, destinations, travel modes, and routes. Comprehensive data to capture travel patterns is today expensive to obtain. Travel surveys can give detailed information by asking travellers questions about their recent travel patterns. Unfortunately, they are very costly to conduct. Therefore, sample sizes are usually small and new a survey is often only conducted every few years. With decreasing response rates, surveys have become even more expensive or less representative (Schulz et al., 2016; Prelipcean et al., 2015). Traffic counts give only the aggregated amount of travellers or vehicles passing at few given locations and no observations of trips from origin to destination.

With technology evolving, there are now several large-scale passive data sources. Here, the data is collected passively, that is, without any additional intervention. These new data sources open new possibilities to obtain large samples of observations of travel patterns. Cellular network data, consists of records of events from mobile phones that cellular network operators capture. We can use these records to obtain large-scale observations of travel patterns with all modes using the cellular network's already existing technical infrastructure. Another large-scale passive data source is smart card data, which is collected by public transportation fare systems. It provides detailed information on historical and current travel patterns with public transportation.

In order to use these new data sources, there are, however, two main challenges. First, the raw data does usually not contain the desired observations directly. Therefore, we need to process the raw data is to extract relevant travel patterns. This calls for new methods capable of handling large amounts of data and distinguishing noise from actual observations. As passive data sources do not always include all metadata needed for analysing travel patterns, new methods are required to infer this additional information. Second, due to the lack of complete ground-truth data, a significant challenge is to evaluate the data processing and the resulting travel patterns to understand their quality.

1.2 Aim and Scope

The aim of this thesis is to propose, compare and evaluate methods of processing large-scale passive data such as cellular network data and smart card data for extracting travel patterns that are relevant for use in traffic planning applications.

The focus of this thesis is on the data processing needed to extract travel patterns. The methods proposed in this thesis aim to extract travel patterns and provide data for traffic planning applications. The large-scale data sources used in this thesis are cellular network data and smart card data.

The area of processing large-scale passive data for analysing travel patterns is broad. Therefore, some delimitations have been made for this thesis. The thesis does not cover details about the data collection of large-scale passive data. The thesis does further not cover the implementation of solutions for specific traffic planning applications. Instead, it focuses on the link between these: the necessary data processing. While two large-scale data sources are used in the thesis, the proposed methods only use one data source at a time. The data fusion of multiple data sources is left for further research. The same holds for the integration with traffic planning models. Finally, the methods only facilitate the analysis of historical data. The adjustments needed to enable real-time processing are not covered in this thesis.

1.3 Methodology

In order to reach the aim of finding and evaluating methods to extract travel patterns from large-scale passive data, we could consider different approaches:

1 *Simulation*: Given artificial travel patterns, we simulate the collection of large-scale passive data. After developing methods for recovering the travel patterns from the simulated data, the methods are evaluated against the artificial travel patterns.

2 *Analytical modelling*: First, assumptions are formulated about the data expected to be collected given certain travel patterns. Then, an analytical method is formulated using the assumptions.

3 *Empirical research*: A dataset of real-world passive data is collected and used to extract travel patterns. The resulting travel patterns are validated using experiments or other available data sources.

Simulation allows to easily control the environment and understand how a method handles data of different quality and resolution. The simulated travel patterns are fully known and can be used for validation. However, it is very difficult to realistically simulate the data collection of, for example, cellular network data which depends on many complex factors such as radio propagation. There is a risk that a method that works well with simulated data performs worse when used on real-world datasets.

Analytical modelling allows obtaining specific quantities using closed form equations, for example, the average distance travelled by a user per day. This methodology is limited to relatively simple quantities since describing comprehensive and complex travel patterns using closed-form equations would be cumbersome if not impossible. Similarly, as for simulation, the method is only useful if the assumptions on data collection hold in reality. Methods solely based on analytical models risk becoming too complex with realistic assumptions or not useful in practice with too simplistic assumptions.

Using a real-world passive dataset to develop and evaluate methods allows testing if a method works under real circumstances. This methodology also allows identifying potential problems and limitations that could be missed in theoretical models or a simulation. A disadvantage of this methodology is that there is usually no complete ground truth for validation. Further, a more complex data processing setup is needed to process real-world, large-scale datasets in a privacy-preserving way.

The main methodology used in this thesis is empirical research: processing and analysing real-world passive data. This methodology allows getting closer to methods that can be used for real traffic planning applications than using simulation or pure mathematical modelling. Three approaches are used to evaluate the proposed methods. The first approach is validation by experiments. Here, data is collected in a way such that the actual travel patterns are known. The second approach is validation by comparison to other data sources. In this approach, the extracted travel patterns are compared to another independent and trusted data source. The third evaluation approach is to compare the travel patterns extracted by different meth-

ods from the same data. This approach provides no validation but can show how sensitive the resulting travel patterns are to changes in the method.

1.4 Outline

The remainder of this thesis is organised as follows. Chapter 2 motivates why the analysis of travel patterns is relevant and introduces the most central terms to describe travel patterns. Chapter 3 gives an overview of different large-scale passive data sources that can provide observations of travel patterns. Both cellular network data and smart card data are presented in detail with their typical characteristics. Chapter 4 then introduces methods to process large-scale passive data for travel pattern analysis. This includes both an introduction to general steps of data processing and a summary of methods that have been used in previous literature to solve problems related to the data processing of cellular network data and smart card data. In Chapter 5 the research questions related to research gaps in the previous literature are formulated. Further, the chapter contains a summary of the research that is part of this thesis and its main contributions. Chapter 6 gives the main conclusions of the thesis. Finally, the thesis contains five research papers.

Chapter 2

Travel Pattern Analysis

Within this thesis *travel patterns* are descriptions of movements of people. Travel pattern analysis aims to understand current, historical and future aggregated travel patterns. This chapter introduces the basic concepts and terms related to travel patterns and illustrates how traffic planners can use travel pattern analysis to improve the traffic system. While later chapters of the thesis focus on using large-scale passive data, this chapter first introduces traffic modelling as a standard approach of travel pattern analysis without using such data.

2.1 Travel Patterns

Travel patterns subsume all relevant information to describe the movements of people. They include information about when, how, where, and possibly why these movements occur in a population. As the term *patterns* suggests, the focus is on describing patterns in the population and not specific travellers. The aggregated patterns are, on the other hand, the result of many individual movements. We can describe travel patterns using the components shown in Figure 2.1. We call travel patterns related to travellers, trips and stops *individual travel patterns* and travel patterns related to the traffic system *aggregated travel patterns*.

Trips are central to describe travel patterns. A trip is a movement of an individual between two stops. The stops correspond to the trip's origin and destination (for example, home and work). A trip also has a start time (departure at the origin) and an end time (arrival at the

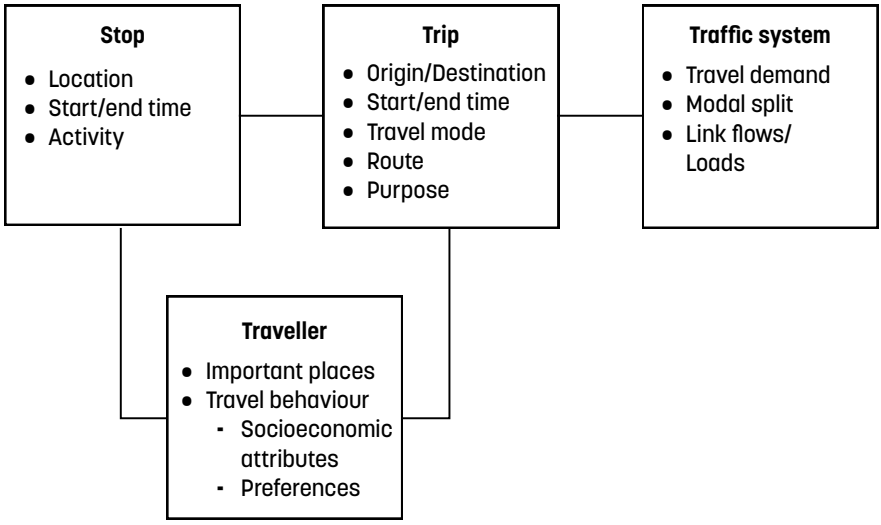


Figure 2.1: The main components of travel patterns and their related attributes.

destination). In addition, we may also associate the trip with the travel mode used, such as private car, train or bicycle. The route of a trip exactly describes which links of the road network or which public transportation lines have been used to make the trip. We may also be interested in a trip's purpose (e.g. commuting, leisure, business, shopping). The purpose can be relevant for traffic planning as it affects the flexibility and the value of time for a trip.

Stops describe the places where individuals are when they are not travelling. They are also called stay-locations. Stops are described by their location, start time and end time. Even the stops are relevant to describe travel patterns because the purpose of trips is to move between stops, and the stops can thus explain why a traveller makes the trip. Therefore, a stop may be associated with an activity category, for example, home, work, leisure, shopping. The activities connected to the stop before and after a trip explain the trip purpose.

The traveller is the individual making the trips and stops. The travel behaviour of the traveller describes the decision-making process when, how, why the individual is travelling. It is influenced by the traveller's socioeconomic attributes, such as income, access to car, age, employment and individual preferences including travel mode preferences. We can also associate some important places with the traveller. These are stops that a traveller frequently visits, for

example, the traveller's home and work.

So far, we have introduced terms related to individual travel patterns. However, it is relevant for traffic planners to understand how individual travel patterns add up to aggregated travel patterns. We can describe the travel patterns on aggregated (macroscopic) level in terms of travel demand. The travel demand is the number of travellers between different areas at a given time. An OD-matrix describes the travel demand by containing the number of travellers (or vehicles) travelling in each pair of *Traffic Analysis Zones* (TAZ). The modal split gives the share of the travel demand made using the different travel modes available in an OD-pair. Finally, link flows describe the number of vehicles (or travellers) using a particular link in the road network. For the public transportation system, we may instead give the loads describing how many travellers use a given line or vehicle on the line. As travel patterns are dynamic, all descriptions of aggregated travel patterns can change over time. Therefore, it is common to give them time-sliced into different time periods.

2.2 Usage and Applications

Suppose that we had a good description of the travel patterns. How can traffic planners make use of it in practical traffic planning applications? We can divide traffic planning into three levels: strategic, tactical and operational planning. An understanding of the present travel patterns is important for all these levels. Strategic planning is about planning from a long-term perspective. That is taking fundamental decisions about developing the traffic system, for instance, by constructing new railroads and roads or investments into an increased fleet of public transportation vehicles. Knowledge of past and current travel patterns is needed to build models that forecast how travel patterns will develop in the future. These models allow traffic planners to evaluate the effect of specific changes or investments on travel patterns and estimate socioeconomic benefits. Making long-term decisions to develop the traffic system also requires a general understanding of travel behaviour. This includes, for example, understanding which factors influence how individuals choose the travel mode to use for a trip.

Tactical planning is about planning the use of the present infrastructure. An example is the development of public transportation

route networks and timetables (Pelletier et al., 2011). An efficient public transportation system needs to adapt when travel patterns change over time. To understand where there is potential to open a new line or extend the timetable, we first need to understand the travel demand. Tactical planning could also include handling planned disruptions, such as a temporary closure of a public transportation route or a road due to construction works. Knowing the present travel patterns allows planning replacements such that the additional travel time is minimised for most people.

Operational planning focuses on short-term decisions. It is also called traffic management. The focus here is on the current traffic situation and handling of unplanned events. Unlike strategic and tactical planning, where we can use historical travel patterns, operational planning requires real-time travel patterns. The real-time data can be used to give adequate traffic information and, for example, reroute travellers in order to minimise queues.

While this thesis focuses on traffic planning applications only, there are also other applications where understanding travel patterns plays an important role. Urban planners may use travel patterns to understand how cities should be developed, for example, to minimise additional traffic generated (Becker et al., 2011b). In cultural geography, travel patterns can be used to understand segregation (Östh et al., 2018). Travel patterns also allow to better understand tourism (Ahas et al., 2007). Finally, travel patterns are crucial for understanding epidemic spread (Barbosa et al., 2018, Section 5).

2.3 Data Collection

Before introducing new large-scale passive data sources in Chapter 3, this chapter gives a brief overview of the main ways of data collection commonly used today to observe and describe travel patterns. One of these data sources are traffic counts. They allow observing the number of travellers or vehicles using specific parts of the traffic infrastructure. Road traffic counts can be used to observe the flow on a specific link in the road network. They can be collected either manually or automatically using temporary or permanent equipment. In public transportation systems, we can collect traffic counts manually or automatically using equipment in vehicles or gates in metro systems. Automatic traffic counts allow collecting updated data fre-

quently. Traffic counts only provide the total number of travellers using a given link or public transportation vehicle. They do not provide any information about where the travellers started their trip or which route they use. Equipment for automatic traffic counts and labour for manual traffic counts is expensive. For this reason, traffic counts are usually only collected at strategic places in the traffic system, such as major roads or places with congestion problems. Therefore, traffic counts provide only partial information about travel patterns.

Travel surveys can, in contrast to traffic counts, provide a sample of individual travel patterns. Participants of a travel survey are asked specific questions about their recent travel patterns. We can obtain metadata for each trip such as the travel mode, purpose and activities before and after the trip by including appropriate questions in the survey. Further, we can collect basic socioeconomic data about the participant and data about personal travel preferences. The knowledge about socioeconomic data also allows to understand and compensate for possible bias in the sample of respondents of the survey.

Unfortunately, travel surveys suffer from decreasing response rates (Schulz et al., 2016). This leads to smaller sample sizes and possibly increasing bias. A problem of self-reported travel surveys is also that there may be underreporting of certain types of trips and inaccuracies in the reported data (Stopher et al., 2007). Recently, efforts are made to replace travel surveys on paper with *Global Positioning System* (GPS) supported travel diaries which could increase data quality and lower costs for carrying out travel surveys (Prelipcean et al., 2015). Even though travel surveys provide a lot of detail about individual travel patterns, the total number of observations included in the sample is usually small in relation to the complexity of the traffic system. A huge sample would be needed to collect observations of all travel modes in all OD-pairs. In practice, this would be too expensive. Due to their cost, travel surveys are also commonly only updated every few years and such that changes in travel patterns are captured with delay only.

Census data and data describing the traffic infrastructure do not provide observations of travel patterns. However, they are still useful to understand travel patterns. Census data can provide the total number of homes and workplaces in an area, and other information on land use that may generate traffic. A detailed description of the available traffic infrastructure is needed to understand, for example, which routes are available. For the road infrastructure, we can asso-

ciate each link with a maximum speed and capacity. For the public transportation system, we need the route network and timetable to understand the exact routes used for trips.

2.4 Traffic Modelling

A small sample of observations is not enough for making well-informed traffic planning decisions. In most cases, we need to have an overview of the aggregated travel patterns of the whole population and in the whole traffic system. Traffic counts and travel surveys provide only samples of travel patterns and not the whole population's aggregated travel patterns. Commonly, traffic models are used to solve this problem and estimate aggregated travel patterns in a population using only limited data.

Traffic models are typically using census data and data about the traffic infrastructure as inputs. The travel patterns are then modelled based on a number of fundamental assumptions. A common assumption made is that travellers seek to minimise their travel time. Traffic models also use parameters that, for example, describe what the experienced value of time is for different groups of travellers. By adjusting the parameters until the resulting travel patterns are in-line with data from a travel survey and traffic counts, we can make sure that the model produces reasonable results. This process is called calibration. By changing the traffic model's input and parameters, it is also possible to compare different scenarios and analyse the effect of changes to the traffic system or in travel behaviour.

A common modelling paradigm used is the four-step model. As the name suggests, it divides the process of modelling travel patterns into four modelling steps: trip generation, traffic demand estimation, mode choice and route choice (see Figure 2.2). In the first step, we model the places generating travel demand using census data about homes, workplaces and other places known to generate traffic. In the trip distribution step we estimate the travel demand: the number of trips induces between the places that generate traffic. Typically, a Gravity model (Erlander and Stewart, 1990) is used for this step. It distributes the travel demand under the assumption that the number of trips in an OD-pair decays with increasing travel time (or generalised cost).

The estimated travel demand is then split among the available

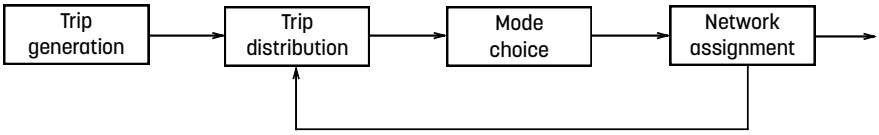


Figure 2.2: The four-step model used in traffic modelling.

travel modes. The travel mode choice is often modelled using a Logit model (Wen and Koppelman, 2001) based on the assumption that most travellers will choose the travel mode that has the highest utility for them by considering factors such as cost and travel time. Finally, in the route choice step, we may use a model that obtains a user equilibrium (Patriksson, 2015). In a user equilibrium, we assign the flow in each OD-pair to the traffic network based on the assumption that each traveller seeks to minimise their travel time. For road traffic, we can model the effects of congestion using fundamental traffic flow theory (Treiber and Kesting, 2013). This allows to estimate the link flows in the network.

The flow assigned to the road links affects travel time in case of congestion. To account for this, we may use an iterative process by using the updated travel times from step four to re-iterate over steps two to four of the four-step model until reaching a stable state. While the methods used in a four-step model are based on assumptions on individual travel behaviour, the travel patterns are only modelled on an aggregated level in terms of flows and not for each traveller individually.

The traditional four-step model is the most common paradigm used in traffic models and modelling software used by agencies and municipalities in practice. However, there are newer paradigms that are considered state of the art by many researchers. In particular, agent-based travel models are seen as an alternative paradigm (Balmer et al., 2009). In these models, we start by modelling individuals and their activities and derive travel patterns from these activities. A strength of these models is that they allow modelling individual and aggregated travel patterns. An agent-based travel model may require even more behavioural data than traditional four-step models to validate the complex individual travel patterns part of the model (Bernhardt, 2007). The simulation of individual agents is also significantly more computationally demanding than models using aggregated flows as in the four-step modelling paradigm.

Travel models have shown to be very useful to describe the overall travel patterns when only limited data from traffic counts and surveys are available. Since the travel patterns are based on the model's assumptions instead of actual observations, the modelled travel patterns do not always agree with the real travel patterns. There are also variations in individual travel behaviour that are difficult to fully capture in a model. Changes in travel behaviour make the model outdated. In developing countries, the lack of adequate data can make it impossible to use traffic models to understand travel patterns.

Chapter 3

Large-Scale Passive Data

With the rise of new technology, new *large-scale passive data sources* sources have emerged that could fill the gaps left by travel surveys and traffic counts. The term *large-scale* here means that the data contains observations of significant parts of the traffic system and not only a small sample, as is the case with travel surveys or traffic counts. *Passive* data means that the data is collected passively, that is, without any additional manual intervention.

This chapter introduces data sources that are both large-scale and passive. These data sources are typically relatively easy to collect as they use existing systems and do not require manual intervention. The large-scale nature allows obtaining large samples of observations. Several large-scale passive data sources are available today. The two data sources covered in this thesis, cellular network data and smart card data, are presented in detail in this chapter with their characteristics, advantages and limitations.

3.1 Cellular Network Data

Cellular network data refers to data that cellular network operators collect for different reasons. Following the taxonomy of different types of cellular network data in Gundlegård (2018), this can include billing data, lcoation updates, handovers, measurement reports and dedicated location data. For the first three types, the user's location is approximated indirectly from the knowledge of which cell the user has been connected to at a particular time. In the case of measure-

ment reports and dedicated location data, a more precise location can be obtained using signal strength and round-trip-time measurements. Data collection efforts and privacy implications are much higher for these types of data. Therefore, measurement reports and dedicated location data are commonly not available to extract travel patterns and are therefore not covered in this thesis.

In the remainder of this thesis, the term cellular network data is used for those datasets that contain billing data, lcoation updates and handovers collected from the cellular network. Billing data is collected when users actively use their phone, e.g., making a phone call or sending a text message. Datasets only based on billing data are also called *Call Detail Records* (CDR). If the dataset includes additional events, the term *x-Detail Records* (xDR) is used in the literature. There are two categories of these additional events: handovers and lcoation updates. Handovers are caused by a switch between cells during an active data connection or phone call (Saifullah et al., 2012). Lcoation updates are recorded, for example, when moving between two location areas. Location areas typically consist of multiple cells and require larger movements to be triggered than handovers. Another type of lcoation updates are periodic updates, which are recorded with a fixed time interval (Calabrese et al., 2014).

The records of cellular network data consist of a user ID, timestamp and the cell ID of the cell to which the user is connected, as in the example in Table 3.1. The frequency of updates and the time resolution depends largely on the type of events. Depending on the antenna density in the area, the estimated position can have spatial uncertainty up to several kilometres (Bhaskaran et al., 2003). A simple but not very accurate approximation of the user's location is the position of the cell tower—the tower on which the antenna of the cell is mounted. A better estimate is to use the coverage area of the cell. We can estimate the coverage using a radio propagation model based on factors such as the antenna's power and the base station's height. However, if such a model is not available, many studies approximate the cell's coverage area using its Voronoi cell (Baert and Seme, 2004). A Voronoi cell for a given cell tower describes the area that includes all positions where this cell tower is the closest (Aggarwal et al., 1989). The Voronoi cells cell can thus be computed only based on the cell tower's locations.

For the analysis of travel patterns, cellular network data can provide large-scale observations of travel patterns with some major ad-

Table 3.1: An example of an artificial cellular network dataset.

User ID	Timestamp	Cell ID
1	2020-10-01 06:50:00	1
1	2020-10-01 08:10:00	3
2	2020-10-01 08:20:00	2
...	...	...

vantages over other data sources. The subscribers of a cellular network operator are typically a significant share of the total population. As mobile devices are ubiquitous today, we can observe movements with all travel modes and follow trips from the origin to the final destination. By using the existing cellular network infrastructure, the effort to collect the data is relatively low. This makes it possible to collect updated data regularly and possibly even in real-time. Further, there is no need to install additional applications on each mobile device.

On the other hand, there are several challenges when using cellular network data. The connection to a specific cell gives only an approximation of the actual position. The accuracy of the estimated position is varying depending on the region. This also means that shorter movements cannot be detected reliably in regions with low cell density. The resolution in time varies for some types of events. In the case of CDR data, the time resolution depends on the user's phone call frequency. Periodic updates, on the other hand, occur with a fixed time interval. Switches between cells can be caused not only by physical movements but also, for example, when the network tries to balance the load between different cells or other effects that influence radio propagation, such as weather conditions. Additionally, phones use different network types such as the *Global System for Mobile communications* (GSM), *Universal Mobile Telecommunications System* (UMTS) and *Long-Term Evolution* (LTE), which can cause additional switches. For analysing travel patterns, these types of switches are noise that needs to be filtered.

When using cellular network data, we need to process the data in a way that protects the privacy of individuals. For this reason, also, cellular network data cannot be linked to socio-economic or other

metadata about the individual users. Therefore, we cannot control potential bias in the sample of individuals in the same way as in travel surveys.

Cellular network data is particularly useful to get an overview of the overall travel patterns in a region as it contains large-scale observations of travel patterns with all travel modes. We may also follow changes in travel patterns over time relatively easy. However, cellular network data only contains the events to observe movements. There is no other metadata directly available on the traveller or the trips. Therefore, data processing is needed to break down the total travel patterns by travel mode, trip purpose (see Chapter 4.3).

3.2 Smart Card Data

Smart cards have been introduced in public transportation systems to increase passengers' convenience and save costs for operators compared to tickets on paper. Besides facilitating ticket purchases, data from smart card systems has also shown to be useful to better understand travel patterns of public transportation passengers (Anda et al., 2017a). Recently, other automatic fare collection methods such as contactless bankcards or smartphone-based ticketing have been introduced (Brakewood and Kocur, 2011). In this thesis, the term smart card data is used, but similar data might be obtained from these alternative systems.

Table 3.2 shows an example of artificial smart card data. In most systems, the data entries contain at least a card ID (which may be rehashed periodically), timestamp and the stop at which the smart card was used. Some systems require passengers to only tap in at the beginning of a trip, while others require passengers to also tap out at the end of a trip. For systems that require to tap in and tap out also the type of the event is recorded. For systems with only tap-ins, the destination of trips is not known and needs to be inferred using behavioural assumptions as discussed in Chapter 4.4. Additional metadata might be available depending on the system, such as the route (line number) used, a vehicle ID or the fare type.

Similar to cellular network data, smart card data is using an existing system and does not require additional infrastructure for the data collection. The data is also easy to update and can potentially be made available in real-time. It can cover a large share of users

Table 3.2: An example of an artificial smart card dataset.

Card ID	Timestamp	Stop	Event type
1	2020-10-01 06:50:00	Central station	Tap-in
1	2020-10-01 08:10:00	City hall	Tap-in
1	2020-10-01 08:20:00	Kings Street	Tap-out
2	2020-10-01 07:20:00	City hall	Tap-in
...	...	...	...

if the smart card system is the main fare system. An advantage of smart card data is that the exact stop and route is recorded directly. Further, we can use the fare type used to understand travel patterns for different groups of passengers.

An obvious limitation of smart card data is that it only covers public transportation. That also means that we have no information on each trip's actual origin and destination, except the first and last stop. In tap-in-only systems, even the last stop used is not known and needs to be inferred. Often smart cards are only one of several parallel fare collection systems. In that case, smart card data might not be perfectly representative of all public transportation users. In general, smart card systems vary a lot between different operators and therefore, we need to adjust the method of extracting travel patterns for the specific system.

Smart card data can be used to improve public transportation systems on strategic, tactical and operational level (Pelletier et al., 2011). Studies on strategic planning, for example, discuss the use of smart card data for planning infrastructure investments, decisions on vehicle investments, forecasting long-term demand, and modelling long-term changes (Briand et al., 2017). Smart card data can also inform tactical decisions, including timetable adjustments, network planning or planning temporary replacement services in case of construction works (Mojica, 2008). On the operational level, we can use it to detect and react to short-term disruptions (accidents, strikes, weather, infrastructure breakdowns), handle large events, provide better traffic information, and monitor performance monitoring (Morency et al., 2007).

3.3 Other Data Sources

Besides cellular network data and smart card data, also GPS traces can be used to analyze travel patterns. GPS tracks can be collected in different ways, for example:

- using a smartphone application or smartwatch continuously in the background,
- when using a smartphone application (for example a navigation app) or
- using devices in vehicles (floating car data).

Advantages of GPS tracks are the high accuracy and possible temporal resolution (Barbosa et al., 2018). The main limitation of the data is its limited availability. Large-scale GPS data owned by companies running the applications is usually not available for researchers and traffic planners. Depending on the specific user group and purpose of the applications collecting GPS data, the data can be heavily biased. It might not represent large parts of the population as, for example, cellular network data does. Thanks to the high resolution, GPS tracks are suitable to estimate travel times and detect traffic states (Hofleitner et al., 2012). Due to the possible bias and incompleteness, it may not be possible to estimate the total travel demand from GPS data. A navigation app, for example, is likely to be used by car drivers for longer routes that are unknown to the driver. It is less likely to be used for everyday commuting trips and may thus underestimate those. Another use case is to conduct travel surveys using an app that collects GPS data (Prelipcean et al., 2015). This, however, requires more user action than other ways of collecting GPS data and might, therefore, no longer classify as passive data.

Several other passive data sources might not provide comprehensive travel patterns as the data sources discussed above but can still be used for specific use cases or complement other data sources. One example is Bluetooth data. We can use Bluetooth sensors along a road to estimate the travel time of vehicles (Haghani et al., 2010; Jaume et al., 2012). For the same purpose, we could also use *Automatic number-plate recognition* (ANPR) (Kazagli and Koutsopoulos, 2013; Cao et al., 2020). Wi-Fi access points can be used to understand pedestrian movements and the number of people at given places (Toch et al., 2018).

Some studies also suggest using social media data, such as geotagged photos or check-in data at places, as a complement to understand travel patterns (Hasan et al., 2013; Cho et al., 2011). While this type of data can provide some insights, for example, which places tourists visit at different times, it is rarely representative for the population and heavily biased due to the user group using the service and the types of captured travel patterns.

Chapter 4

Methods for Processing Large-Scale Passive Data

Large-scale data sources may provide large amounts of observations. However, these observations are not directly containing the comprehensive descriptions of travel patterns needed for traffic planning. The data often contains noise and lacks the necessary metadata needed to perform a travel pattern analysis. Therefore, using large-scale passive data sources to analyse travel patterns requires extensive data processing to extract the travel patterns from the raw data.

This chapter introduces the typical steps to analyse travel patterns from large-scale passive data and common data processing methods for extracting travel patterns from large-scale passive data. An overview of how these data processing methods have been used in previous literature for typical problems when using cellular network data and smart card data for travel pattern analysis is given. Several approaches are discussed to evaluate the data processing methods with respect to the quality of the extracted travel patterns. Finally, different ways of using the resulting travel patterns for traffic planning applications are discussed.

4.1 Steps for Processing Large-Scale Data

Independent of the data source and application, we need to consider some general steps when analysing travel patterns from large-scale

passive data (see Figure 4.1). A typical analysis starts with the data collection. Then, we use data processing to extract the travel patterns required for the application. This step often starts by cleaning the data followed by several processing steps. Next, we evaluate the resulting travel patterns to understand if the method for extracting the travel patterns works as required. Finally, we may use the extracted travel patterns for real traffic planning applications.

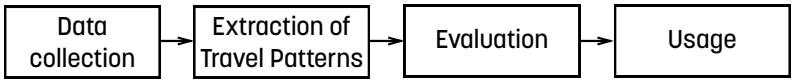


Figure 4.1: Typical steps to analyse travel patterns from large-scale passive data.

It is important to know how the data is collected to select appropriate data processing methods and interpret the analysis results correctly. We should be aware of possible bias caused by the way the data is collected. Cellular network data collected by an operator with a very special customer group that is not representative of the population could, for example, lead to an overrepresentation of particular travel patterns. Before choosing the processing method, it is crucial to understand the data's characteristics. This includes the general characteristics of the data source as described in Chapter 3, but also the characteristics of the specific dataset, for example, its resolution in space and time. It is often required to at least reconfigure the method's parameters to apply the same method to a new dataset that has been collected slightly differently. It is also common that the data is manipulated to ensure the privacy of individuals. Examples are the periodic re-hashing of user identifiers and the obfuscation in time and space (see Chapter 4.6). The data processing method needs to take into account how the data has been manipulated.

After collecting the raw data, we need to process it to extract travel patterns. A large-scale dataset often contains incorrect data points or data that is not relevant for the particular analysis. Therefore, data processing usually needs to starts with some kind of data cleaning. The data cleaning aims to identify possible problems in the data. First, we should analyse how much of the data is affected by the problem. The data cleaning should then try to remove such data that is obviously incorrect. We may also filter out data that is irrelevant to the application. If the use case is limited to a particular

region or time period, we may filter the raw data for observations in that region and time period only.

After data cleaning, the data processing continues to extract the travel patterns relevant to the application. The method for extracting the travel patterns needs to consider both the data characteristics and the requirements of the analysis. The extraction of the travel patterns may be divided into several steps, for example, a trip extraction step followed by a travel mode classification step. Relevant processing steps and related methods for cellular network data and smart card data are given in Chapters 4.3 and 4.4. The data processing needs to be designed and implemented in a computationally efficient way to process large-scale data within a reasonable time. When designing a method, we can use computational complexity to compare different algorithms (Hartmanis and Stearns, 1965). Two common tools to process large-scale data efficiently are the concept of MapReduce and the use of database management systems (Pavlo et al., 2009). MapReduce divides the data into chunks that are processed in parallel and then joined together and reduced into the desired result. Database management systems reduce computation time to query for specific data using indices and efficient data storage.

After implementing the data processing method, we need to evaluate it. The evaluation aims to ensure that the method is working correctly and to understand the extracted travel patterns' quality. Different evaluation methods are to collect data in a controlled experiment, compare with other data sources on an aggregated level, and compare the result of applying different data processing methods. These methods of evaluation are discussed in Chapter 4.5.

After the data processing and evaluation, we can use the travel patterns for a particular traffic planning use case. The travel patterns can be analysed directly, for example, using statistical methods and visualization. We may also combine them with other data sources or traffic models. Different ways of using the extracted travel patterns are presented in Chapter 4.7.

4.2 Data Processing Methods

Several data processing steps are often needed to process the raw data to gain the information relevant for a particular application. Common types of data processing steps are extraction, aggregation

and estimation, inference and classification. An extraction step has the purpose of filtering the data for relevant parts; an example is the extraction of trips from cellular network data. Aggregation and estimation steps are about estimating quantities based on the data—for example, the estimation of travel demand. Inference steps aim to make conclusions based on the data, for example, re-identifying the most likely route on the road network used for a trip. Classification is a variant of inference that aims to place observations into different given categories, such as classifying the travel mode used.

We implement each data processing step using a data processing method. We can group data processing methods into rule-based algorithms on the one hand and machine learning methods on the other hand (see Figure 4.2). Rule-based methods are using heuristic algorithms or explicit queries that filter data for certain criteria. Rule-based methods are usually based on behavioural assumptions (for example, that most people choose the fasted route alternative) and assumptions about the data collection (for example, that the probability of connecting to a cell in the cellular network decays with the distance from that cell). Rule-based methods can involve many parameters and thresholds, which we can set empirically or using systematic calibration. It is also common in rule-based methods to use other data, such as geospatial data about the transportation network.

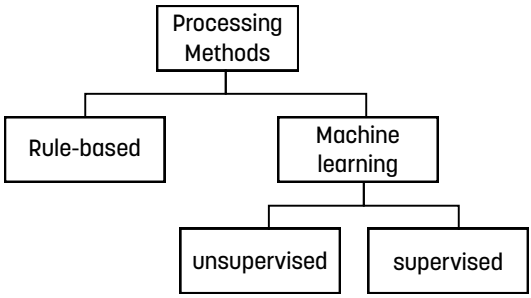


Figure 4.2: Taxonomy of data processing methods.

Machine learning methods use a different approach. Instead of formulating explicit rules and thresholds, machine learning methods automatically identify patterns. Compared to rule-based methods, machine learning methods typically use fewer explicit assumptions and parameters. Machine learning methods also tend to perform better with an increasing amount of data available for learning, which is not the case for rule-based methods. Common categories of machine

learning are supervised learning and unsupervised learning (Toch et al., 2018). In supervised learning, we use a training dataset with the correct result known for the learning process. After training, we can apply the method to predict the result for a new unseen dataset. Unsupervised learning methods try to find structures in a dataset without any training data.

Classification problems can be solved using supervised learning. We train a classification method using training data for which the correct class label is known. Other supervised methods include regression models, which we can use to infer quantitive outputs rather than categories (Liero and Zwanzig, 2016). A disadvantage of supervised methods is that enough training data with known output needs to be available, which often is expensive to collect.

Clustering is an unsupervised learning problem (James et al., 2013). Instead of a given set of categories as it is the case in a classification problem, clustering only uses unlabeled data to define groups (clusters of observations) in the data such that the observations in each cluster are similar. Dimensionality reduction is another way of unsupervised learning. It allows reducing the number of properties (features) to describe an observation. A popular method of dimensionality reduction is *Principal component analysis* (PCA) (Wall et al., 2003). Some learning methods can be used for both supervised learning and unsupervised learning. Examples are *Hidden Markov Models* (HMM) and neural networks (Rabiner and Juang, 1986; Anderson, 1995).

Whether the use of a rule-based or machine learning is most appropriate depends on the problem. An advantage of rule-based methods is that they can usually be understood more intuitively than machine learning methods and allow in particular to investigate exactly how observations are processed. A difficulty of rule-based methods is making correct assumptions and finding good parameter values. Especially in the absence of ground-truth data or when there are many parameters, systematic calibration might not be feasible. When the patterns to detect are complex and challenging to describe with manual rules and parameters, machine learning may be more appropriate. A disadvantage of using machine learning methods is that it is more difficult to understand why particular observations have led to a particular result.

4.3 Cellular Network Data Processing

As cellular networks have not been designed as a positioning system, the data is often noisy, of low resolution in time and space and lacks additional metadata. Therefore, data processing is necessary to extract travel patterns from cellular network data. Common problems discussed in the literature are the extraction of trips, estimation of travel demand, mode classification, route inference and trip purpose and activity inference (Anda et al., 2017b; Wang et al., 2018). Methods to solve each of these problems are discussed in this chapter.

As many applications require several of these processing steps, we may organise them in a pipeline of different data processing steps executed sequentially or parallelly. Suppose the application is to estimate link flows on the transportation network. As cellular network data can contain updates even when a user is not moving, the processing needs to start with a trip extraction step. It is then necessary to estimate the total travel demand and separate the extracted trips by mode to associate them with the proper infrastructure. A final data processing step could then infer the routes in the transportation network used to load the flows on the transportation network and calculate aggregated link flows. We can execute these steps in a data processing pipeline as shown in Figure 4.3.

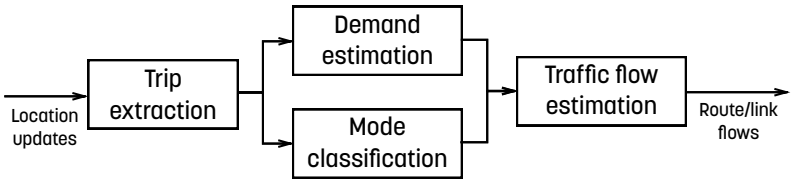


Figure 4.3: Example of a data processing pipeline to extract link flows from cellular network data.

4.3.1 Data Cleaning and Trip Extraction

As the first step of data processing, several studies use a data cleaning step to remove noise from the raw data (Wang et al., 2018; Alexander et al., 2015; Huang et al., 2019). A common type of noise in cellular network data is the oscillation between cells ("ping-pong events"). Heuristic rules can be used to detect and remove these patterns (Wu et al., 2014). We can use similar rules to remove other outliers and

errors in the data, such as, for example, unreasonable large or fast hops between cells.

To analyse travel patterns, usually, only periods of movement are of interest. To identify these periods from cellular network data, a trip extraction step is used in most studies (Wang et al., 2018). We can describe each trip by its start and end time, the origin cell and destination cell and its cellpath. The cellpath is the list of updates recorded during the trip (containing the cell ID and timestamp of each update). The trip extraction method has to consider that cell switches can be made without physical movement, noise and errors in the data, and the limited time and spatial resolution of the cellular network data.

Several methods to extract trips have been proposed in the literature. We can group them into three main categories: frequency-based trip extraction, stop based trip extraction and movement-based trip extraction. Frequency-based trip extraction is first extracting the most frequently visited locations of a user, which for example, correspond to home and work (Alexander et al., 2015; Gundlegård et al., 2016; Isaacman et al., 2011). These locations can be found by querying for the cells that a user most frequently connected to or using a clustering method. A trip is detected when an update occurs at a different location beyond some distance threshold from the last visited frequent location. This threshold is necessary since otherwise, noise not related to real movements would be extracted. This method is handy for sparse CDR data and for extracting commuting trips. It allows inferring the possible origin and destination of a trip even when the data is incomplete using behavioural assumptions such as that users always start and end their day at the home location.

If the data contains more frequent updates than CDR data, stop or movement-based trip extraction can be used. Movement-based trip extraction aims to directly identify periods of continuous movement, for example, using speed and direction (Breyer et al., 2017). Stop based trip extraction instead focuses on detecting stops (stay locations) in the data and then defines periods between stops as a trip. Stop based methods can be implemented using rule-based algorithms, for example, using a threshold for distance and duration that a stop needs to fulfil. Other authors propose to use spatio-temporal clustering to identify stop locations (Gonzalez et al., 2008; Toole et al., 2015). Stop based trip extraction is the most common method used for trip extraction in the literature (Calabrese et al., 2011, 2010; Ming-Heng

et al., 2013; Bachir et al., 2019; Breyer et al., 2017).

4.3.2 Travel Demand Estimation

The goal of travel demand estimation is to estimate the number of travellers between different areas, typically described in an OD-matrix. At first sight, the estimation of travel demand could be done by simply aggregating previously extracted trips. However, there are mainly two reasons that require additional processing: The first is that trips extracted from cellular network data of one operator do not cover the whole population. For this reason, scaling is required. The second reason is that the extracted trips are usually described by their origin and destination in terms of a cell in the cellular network. However, for practical traffic planning applications, the travel demand should be converted to appropriate TAZs instead.

The simplest scaling method is to multiply all trips with a constant factor based on the number of customers of the operator in relation to the population. However, this will not compensate for any bias in the extracted trips. Different types of inherent bias may occur when extracting trips from cellular network data (Chen et al., 2016). For example, there may be operator bias (different operators have different customer groups), regional bias (different operators might be more or less represented in different regions and mobile usage bias (users that use their device more frequently can generate more events). The characteristics of the data may cause further bias. An example is trip length bias caused by the fact that longer trips are detected more reliably than shorter trips.

In travel surveys, we can control some bias by making sure that the participants' composition is representative of the population with respect to socioeconomic attributes. For cellular network data, we cannot use this approach as we have no socioeconomic attributes linked to individuals. However, several scholars suggest scaling methods using more than just one scaling factor for all observations (Calabrese et al., 2013). One method is, for example, to use separate scaling factors for different geographical zones. Alexander et al. (2015) scale trips from cellular network data using the number of mobile users with an estimated home location in a zone relative to the population in the zone according to the census.

The conversion to TAZs is needed to use the estimated travel demand for traffic planning applications. It is also needed to be able

to compare or combine an estimated OD-matrix from cellular network data with another OD-matrix which, for example, has been estimated from a model using travel survey data. Here, a simple method is to assign each trip to the origin-destination pair corresponding to the zones containing the trip's origin and destination cell tower. However, given that cells can have large coverage areas, the cell tower's position may not be a sufficient proxy, especially if the TAZs are not much larger than the cells. In that case, a cell might overlap with several TAZs, and it might thus be difficult to assign each trip to exactly one OD-pair of TAZs. An approach to solve this is splitting the trip and assigning "fractions" of the trip to all relevant OD-pairs. In general, it is easier to estimate travel demand for large TAZs than very small TAZs from cellular network data as found by Batran et al. (2018).

4.3.3 Travel Mode Classification

Many traffic planning applications also require understanding how the travel patterns are split among different travel modes. This is naturally the case for all analysis related to mode choice and modal split estimation. However, it is also relevant when estimating link flows in the transportation network based on the estimated travel demand since the flows need to be assigned to the infrastructure that belongs to the chosen travel mode. The mode classification problem is to label trips extracted from cellular network data by travel mode. The travel modes used for classification vary among different studies. Many authors focus on modes that use different infrastructure (rail, road, air) since these are easier to detect from sparse cellular network data. Only a few studies try to detect more fine-grained modes such as bus, car, tram (Huang et al., 2019).

The processing methods discussed in the literature to classify travel mode use different approaches, including both rule-based and machine learning methods (see Chapter 4.2). A rule-based method used by Kalatian and Shafahi (2016) is to use the characteristics of a trip, such as the travel speed, to classify travel mode. While this is an intuitive method, it may not be possible to estimate the actual speed accurately if the used cellular network data does not contain frequent updates. Also, the speed of several modes may often be too similar to classify the travel mode certainly. Another rule-based approach is to consider other geometric data such as the transportation network or available route alternatives (Qu et al., 2015; Phithakkitnukoon et al.,

2017). Here, we predict the travel mode with the most similar geometry by comparing the trip's cellpath to the infrastructure or route alternatives. This method requires that the used route is found as an alternative and can fail if the route alternatives for several modes are very close.

Machine learning methods for mode classification include both supervised and semi-supervised methods. Supervised methods require labelled training data, that is, trips that have been manually labelled by the correct travel mode (Xu et al., 2011). Supervised methods can perform better than rule-based methods by using patterns such as specific cell sequences that are common when travelling with a given mode. Enough labelled training trips are needed to allow the method to learn all relevant patterns and not overfit. This training data also needs to be available for each OD-pair to learn patterns specific to different OD-pairs. As obtaining manually labelled data is expensive, this is usually not feasible. Semi-supervised approaches (van Engelen and Hoos, 2019) are an alternative where only small amounts of labelled data are required since also patterns in unlabeled data are used. Bachir et al. (2018) for example, use a cluster-then-label method to classify trips by travel mode.

4.3.4 Route inference and Traffic Flow Estimation

An important application in traffic planning is the estimation of traffic flows in the transportation network, which allows us to understand how the network is used and where capacity is exceeded. Route inference aims to find the most likely route taken in the transportation network for a given trip. The traffic flow estimation is about estimating aggregated route and link flows in the transportation network. Several studies have proposed methods to infer the used routes for trips detected from cellular network data and the estimation of traffic flows.

A straightforward route inference method is to calculate the shortest or fastest path in the road or public transportation network, as discussed by Gundlegård et al. (2016); Kanasugi et al. (2013). However, this method is based on the strong assumption that all travellers always choose the shortest path, which is not always true. To handle situations where there are multiple reasonable route alternatives to choose from, some authors propose to see the route inference problem as a classification problem instead with the aim to select

the most likely route used among a number of available alternatives (Becker et al., 2011a). Other methods try to use the available information about a trip, such as its cellpath, to reconstruct the most likely route used using map-matching techniques (Fillekes, 2014; Dash et al., 2015; Algizawy et al., 2017; Song et al., 2017). Here, either rule-based strategies such as the shortest path calculation using modified link costs as proposed by Leontiadis et al. (2014) or machine learning methods such as HMM models are proposed (Jagadeesh and Srikanthan, 2017). Instead of inferring a route for every trip, some authors use a route choice model (Schlaich, 2010) or traffic assignment method (Tettamanti et al., 2012; Huang et al., 2018) to obtain aggregated flows. The observed trips from cellular network data are then used as input or for calibration.

One method to estimate traffic flows proposed by Gundlegård et al. (2016) is to infer routes in the transportation network for all trips and count how often each link has been used. However, this requires that route inference can be made reliably for all trips. This might not be possible if the transportation network is very dense and the cellular network data sparse, as in the case of CDRs. Further, estimating absolute traffic flows of vehicles on the road network requires scaling the trips from cellular network data, typically corresponding to persons, to the number of vehicles. An alternative method is to use a model that considers the activity of mobile phones when estimating flows without inferring an exact route for individual trips, as suggested by Tettamanti et al. (2012). The development of data-driven traffic models, as described by Tsanakas et al. (2021), could allow creating a model to estimate traffic flows that can use cellular network data.

4.3.5 Trip Purpose and Activity Classification

Some traffic planning applications require knowledge about the purpose of a trip and the types of activities between which the trip was made. A commuting trip to work, for example, may have different time constraints and value of time than a leisure trip. While cellular network data does not allow to get detailed insights about activities as in a travel survey, several attempts have been made to enrich trips extracted from cellular network data with metadata about the trip purpose. In the absence of definitive knowledge, we can use several features to classify trips by purpose. These include patterns of time,

duration, frequency and regularity of visiting a place (Dong et al., 2015). Additionally, some authors also consider data on land-use (Dong et al., 2015) and transition probabilities between activities to aid the classification of activities (Yang et al., 2016).

The methods discussed in previous literature contain rule-based and machine-learning methods even for the data processing for trip purpose and activity classification. Frequency-based trip extraction methods, as discussed in Chapter 4.3.1, use frequently visited places such as home and work. This gives already a simplistic rule-based classification of stop activities (Isaacman et al., 2011; Toole et al., 2015). Rule-based methods usually rely on time and frequency rules defined for different types of importation locations that correspond to activities. A problem with these methods is that the time and frequency rules may not hold for all groups of people (consider, for example, night-shift workers or non-commuters). Defining rules for many different activities manually does not scale well and involves many assumptions that are difficult to verify.

Machine learning methods include clustering techniques and probabilistic models (Anda et al., 2017b). Dong et al. (2015), for example, use a clustering-based method. Among probabilistic models, for example, the use of a relational Markov network is proposed by Widhalm et al. (2015). Yin et al. (2017) are instead using a generative model that tries to learn activity chains. An advantage of a generative model is that we may use it to generate simulated activity chains, which is useful for agent-based traffic models (Balmer et al., 2009).

4.4 Smart Card Data Processing

Smart card data in public transportation systems is only collected when a user is travelling. Therefore, it is unnecessary to identify when a traveller is making a trip in the same way as for cellular network data. However, in some systems, destination inference may be needed as a first data processing step to extract complete trips. Other common data processing steps resemble those for cellular network data and include route inference, travel demand estimation, load estimation, trip purpose and activity classification. Some studies also try to segment public transport users into different groups of travellers.

Destination inference is relevant for systems that only provide tap-in data for each trip and do not require the user to tap out at the

end of the trip (see Chapter 3.2). It is widespread to use rule-based heuristics for destination inference. A common heuristic is to use the origin of the next trip as the destination of the previous trip (He and Trépanier, 2015; Trépanier et al., 2007). It is also possible to use machine learning methods for destination inference. Jung and Sohn (2017) for example, suggest a method based on deep learning. We can also use a route inference method to reconstruct the route and itinerary of the trip if the system does not provide this information directly (Zhao et al., 2017; Chu and Chapleau, 2008).

Similar to cellular network data, smart card data may be used to estimate travel demand based on the (reconstructed) trips (Munizaga and Palma, 2012; Alsger et al., 2015). In contrast to cellular network data, smart card data can by design only provide an OD-matrix of trips made using public transportation. The OD-matrix will also be between stops in the public transportation system rather than the actual origin and destination, including possible access modes. It is possible to aggregate the OD-matrix to TAZs if needed.

We may also classify trip purposes and activities from smart card data. The proposed data processing methods for this problem are similar to those for cellular network data. We can infer home and work locations using rule-based approaches, as proposed by Sari Aslam et al. (2019). To infer more detailed trip purposes, Alsger et al. (2018) propose a model that involves travel survey data and land use data in addition to smart card data. Han and Sohn (2016) propose an HMM based method.

Another problem discussed in the literature is the load estimation using smart card data. This problem corresponds to the estimation of traffic flows (route or link flows) for road traffic using private vehicles. The goal of load estimation is to estimate the number of travellers on a given line or even individual vehicles on the line (Chu and Chapleau, 2008).

Apart from classifying activities connect to trips, it may also be useful to segment users of a public transportation system into different groups. This allows analysing travel patterns per passenger group. One segmentation method is to identify predefined groups of interest, such as, for example, commuters (Ma et al., 2017) using patterns such as temporal and spatial regularity. To perform an explorative analysis, if the groups are not known, clustering methods can segment passengers into groups with similar travel behaviour. Several studies have used k-means clustering or hierarchical clustering methods to

group passengers by their temporal patterns (Viallard et al., 2019; Deschaintres et al., 2019; Egu and Bonnel, 2020). Some authors also include spatial patterns such as the regular use of the same route to group passengers (Manley et al., 2018; Morency et al., 2007).

4.5 Evaluation of Data Processing Methods

Before using a method for processing large-scale passive data to extract travel patterns for real applications, we need to evaluate the method. The evaluation of the method has two purposes. First, to validate the method and understand the extracted travel patterns' quality. Second, to understand if shortcomings in the extracted travel patterns are likely an inherent limitation of the data or a particular method's limitation. The following questions are of interest to evaluate a method:

1 Are the extracted travel patterns correct for individual travellers?

2 Are the extracted travel patterns in line with the expected travel patterns in the population?

3 How sensitive are the extracted travel patterns to changes in the method or its parameters?

The first question asks for one type of validation: the validation of the travel patterns extracted by the method on an individual level. We can usually not address this question only using a large-scale dataset. The reasons are that inspecting individual data is usually not possible due to privacy implications and that the correct travel patterns for each individual are unknown. Instead, we may best address this question using experiments: the controlled data collection such that the correct travel patterns are known. For that purpose, data is collected in the same way as large-scale data by a smaller number of individuals who have given their consent. The participants may either annotate their travel patterns manually or aided by another independent technology. An example is to collect both cellular network data and GPS tracks on all participant's devices (Fillekes, 2014). This parallel data collection establishes a ground truth that we can use for validation.

Typical metrics for validation are recall and precision. Recall gives the share of observations in the ground-truth data that have also been extracted by the method. In contrast, precision gives the share of observations that have been extracted by the method and have a matching observation in the ground-truth data. For classification problems, a test error can be calculated, giving the share of correctly classified observations. An example of this method of validation is found in Chin et al. (2019) who have used manually tagged trips in order to calculate recall and precision of a mode classification method.

The second question asks for another type of validation: the match of the extracted travel patterns and the travel patterns in the population. Here the focus is on the travel patterns on an aggregated level. We can address this question by comparing the resulting travel patterns with other data sources. We use the method subject to validation to extract travel patterns from a large-scale dataset. Then, we compare the results to another independent and trusted data source. This data source may for example be a travel survey (Liu et al., 2014; Batran et al., 2018; Bachir et al., 2019), the output of an existing traffic model such as an OD-matrix or traffic flows (Yin et al., 2017), traffic counts from road sections or public transportation (Wismans et al., 2018) or census data (Vanhoof et al., 2018).

To compare to another data source, it is often necessary to convert and/or aggregate the travel patterns to make them comparable. To get an overview, we can compare summary statistics such as the total number of trips or the modal split. Typical metrics to compare with another data source include correlation metrics such as R^2 or statistical tests. Finding a good comparison metric can be challenging for complex spatio-temporal travel patterns. Generic correlation metrics may not always be the best way to measure similarity. Specialised metrics may be more appropriate as discussed by Pollard et al. (2013) for the comparison of OD-matrices and Balakrishna et al. (2015) for the comparison of link flows.

The third question is asking about the output's sensitivity to changes in the method. Answering this question is not enough to validate a method. However, it can help understand if potential shortcomings in the extracted travel patterns are a limitation of the particular method or the data itself. A comparison of methods can provide answers to this question. First, we apply different methods or the same method but with varying parameter values to the same dataset. Then, we compare the resulting travel patterns to measure

their similarity and possible systematic differences. In general, the comparison of methods can be made on an aggregated level, for example, by comparing the modal split (Wang et al., 2010). However, it can also be made on an individual level as done by Vanhoof et al. (2018) who compare the estimated home location for each user estimated by different methods.

The comparison can be made in similar ways as the comparison with other data, for example, by calculating the correlation between different runs. Comparing different methods or parameter setups cannot be used to conclude on the extracted travel patterns' quality. In some cases, however, plausibility indicators can measure if intuitively expected assumptions are fulfilled in the resulting travel patterns. A reasonable assumption is, for example, that when travellers travel back and forth between two cities on the same day, usually, they use the same travel mode for both directions. To evaluate a travel mode classification method, we may thus calculate the share of users who use the same mode for return journeys as a plausibility indicator.

The three evaluation methods fulfil different aims which complement each other (see Table 4.1). Experiments can show that a method can extract travel patterns correctly. However, the significance of the quantitative results largely depends on the size and representativeness of the collected data. Since the data collection in experiments requires much effort, it is usually not feasible to collect data representative of the population. Experiments may thus only give indications of the performance of the method on a large-scale dataset. The comparison to another dataset may be used to judge if the extracted travel patterns' quality fits the application. It can reveal whether there are systemic errors or bias in the extracted travel patterns. A common issue is that the available data is not quite comparable as it is, for example, from a different year (Calabrese et al., 2014). Using very aggregated data, such as an aggregated modal split from a travel survey, does not allow us to understand the cause of potential errors. The aggregation level also affects the comparison result (Batran et al., 2018). Often the data source used for comparison makes no perfect ground-truth either, which is, in particular, true for the output of models. A major advantage of the comparison of methods is that no other data is required. This evaluation method can give insights on sensitivity and systematic differences when using different extraction methods. However, the comparison of methods cannot replace experiments and the comparison with other data which are needed in order

to judge the quality of the resulting travel patterns.

Table 4.1: Comparison between evaluation methods for evaluating methods for extracting travel patterns from large-scale passive data.

	Experiments	**Comparison with other data**	**Comparison of methods**
Aim	Validate that the method correctly extracts individual travel patterns	Validate quality and find systematic errors	Evaluate sensitivity to changes in the method
Comparison data	Actively collected data with known individual travel patterns	Independent data source or model output	Results from running different methods on the data
Typical metrics	Recall/Precision, Test error, Time/spatial error	Correlation, Statistical tests, Summary statistics	Correlation, Statistical tests, Summary statistics, Plausibility indicators

How different evaluation methods complement each other can be illustrated, for example, in the case of a trip extraction method for cellular network data. One way of evaluation is to compare trip-by-trip using experimentally collected data. As ground truth, we may use another data collection method such as a GPS based travel diary (Fillekes, 2014). This allows us to understand if the method works correctly but cannot fully answer if it works in a large-scale setting. For this, we may complement the evaluation by a comparison with other data such as aggregated statistics on the distribution of trip lengths, the number of trips per day etc., since they can be obtained, for example, from a travel survey (Liu et al., 2014). By testing various methods and parameters, we may get additional insights about whether potential limitations in which trips are extracted are due to flaws in the method or inherent limitations of the data.

4.6 Data Privacy

When processing large-scale observations of human mobility, a major concern is to maintain the data privacy of individuals. For traffic planning applications, travel patterns in the whole population and not for specific individual travellers are of interest. However, the large-scale passive data sources presented in Chapter 3 contain observations related to individuals. Even though the observations are usually pseudonymised, there is still a risk that individuals may be re-identified in the raw data and possibly reveal their movements. The data processing of large-scale passive data, as described in the previous sections, therefore, needs to be done in a privacy-preserving way.

De Montjoye et al. (2018) describe four approaches used in the literature to protect privacy while still being able to use the data to extract anonymised and aggregated data, such as travel patterns. The first approach is *Limited release*. In this approach, the raw data is transformed, for example, using periodic re-hashing of user-ids to prevent tracking over time (Toch et al., 2018). Other methods include sampling a subset of the observations or obfuscating timestamps and locations by adding random noise. The transformed data is then shared with trusted partners. An advantage of this approach is that it does not require a complex setup at the data provider and allows researchers to explore the data more directly than in other approaches. This makes the approach particularly useful for initial exploratory analysis. A downside is that the data is leaving the data provider. Studies have shown that there may still be a relatively high risk of re-identification of individuals if the data has not been manipulated enough (Zang and Bolot, 2011). If the data is transformed too much, on the other hand, its utility decreases.

A second approach described by de Montjoye et al. (2018) is *Remote access*. Here, the raw data stays in control of the data provider (the operator in case of cellular network data). Only high-level aggregated data without connection to individuals is being exported and shared for later analysis. The idea of this approach is sometimes described as "bring the code to the data" (Kaisler et al., 2013), as opposed to the usual "bring the data to the code". The clear advantage of using remote access is that the data never leaves the data provider's premises, ensuring privacy through security. It is possible to use a setup where the code is executed fully automatically and only

aggregated results are exported. This also facilitates processing data in real-time. The challenge of this approach is that the data provider needs to maintain a rather complex setup.

A third approach is using *pre-computed indicators and synthetic data* (de Montjoye et al., 2018). The data provider does not share the actual data but indicators based on the data in this approach. An example could be aggregated data without connection to individuals, such as the total number of switches between cell towers. It could also be aggregated statistics about the data. These could allow generating a synthetic dataset using simulation which resembles the original data. While this approach ensures privacy well and may even permit to release data openly, much information may be lost in the process, which significantly limits the possible use cases.

The fourth approach described by de Montjoye et al. (2018) is *Question-and-answer*. Here the data provider provides an *Application Programming Interface* (API) that third parties can use to submit queries. Even in this approach, the real data does not leave the data provider, and only aggregated results that fulfil some anonymity criteria are shared. This approach requires the most complex setup on the data provider side, including the data processing abilities to answer the queries. On the other hand, it allows to specify which data can be exported and permits logging of all queries.

Pre-computed indicators and synthetic data, as well as Question-and-answer approaches, are suitable for production solutions. They are, however, not suitable for the research on methods for extracting travel patterns from large-scale passive data since the data shared in these approaches has already been processed. Therefore, Limited release and Remote access are most appropriate for developing new methods for the extraction of travel patterns. Later on, however, these methods can enable the processing behind Pre-computed indicators and synthetic data and Question-and-answer solutions that may be used for practical applications and analysis.

4.7 Usage of the Extracted Travel Patterns

The focus of this thesis is on the methods to extract travel patterns from large-scale passive data. However, in this chapter, an overview shall be given about how the extracted travel patterns can be used in traffic planning applications. The three main approaches presented

here are the direct analysis of the data, the data fusion with other data sources and the integration with traffic models.

Due to their large-scale nature, it is possible to directly analyse travel patterns extracted from cellular network data and smart card data without using additional traffic models. Quantities that can be estimated include travel demand, mode choice, modal split and traffic flow, among others (Caceres et al., 2008). Map visualisations, for example, allow identifying where there is congestion and where the development of the traffic infrastructure makes sense. A use case of smart card data is also the supervision of the public transportation system's performance, including load monitoring, estimating transfer times, and punctuality (Morency et al., 2007).

It is also possible to obtain general statistics on travel behaviour based on the extracted travel patterns. Examples are the distribution of trip lengths, average number of trips per individual or the number and displacement of locations visited by individuals (Alessandretti et al., 2020). Since the data is easier to update than a survey, we can use it to quantify changes in travel patterns after a change in the traffic system, such as opening new infrastructure or a change of the fare system for public transportation. If the data does not cover the whole population, scaling is needed to draw conclusions about the population's aggregated travel patterns. The estimation of total travel demand from cellular network data of one operator, for example, requires scaling the detected trips in order to match the total population. For this, another data source is needed, for example, census data.

Several tools may be helpful to enable traffic planners to draw conclusions from the extracted travel patterns. Since the extracted travel patterns usually are complex, further aggregation may be useful to get the bigger picture. Depending on the purpose of the analysis, we may perform this aggregation in time or space. Sometimes it may also be useful to aggregate by different groups of travellers to identify the difference between groups. The aggregation can be done manually, for example, by grouping all trips by weekdays. It is also possible to use clustering (see Chapter 4.2) to automatically identify groups for aggregation. An example is identifying different groups of travellers by their travel behaviour (Deschaintres et al., 2019). An essential tool to understand and draw conclusions from the data is visualisation. Two examples are plots of time series or distributions (histograms) of values and alluvial diagrams. Alluvial diagrams are

a particular type of diagrams used to visualise flows. To visualise the spatial dimension of travel patterns, using a *Geographic Information System* (GIS) is useful to create maps of traffic flows or travel demand.

While the direct analysis of travel patterns extracted from large-scale passive data is possible for some applications, there are limitations. One limitation may be the quality and completeness of the extracted travel patterns due to the raw data's limited resolution and possible bias. We can address this problem using data fusion. The idea of data fusion is to combine the extracted travel patterns from large-scale passive data with other data sources. This may both include travel patterns from a second large-scale passive data source but also other data sources. An example for example to combine travel patterns from cellular network data and smart card data. Cellular network data is especially useful to estimate the total travel demand, while it is difficult to estimate how many travellers use each travel mode in many cases. Combining the cellular network data with smart card data a better estimate of the number of total travellers and those travelling by public transportation could be obtained. Bwambale et al. (2017) propose to combine travel patterns extracted from smart card data with travel survey data to improve the understanding of travel behaviour. Wu et al. (2015) suggest combining travel patterns from cellular network data and traffic sensor data to estimate road flows.

The data fusion of multiple data sources can increase the quality and, where data sources complement each other, add more detail to the travel patterns. As data sources may have different formats and aggregations levels, a conversion may be necessary. We can consider optimisation methods to find travel patterns that match as much as possible with both data sources. The data fusion method needs to be designed specifically for the data sources to combine. The result of the data fusion can then be used in the same way as directly analysing travel patterns extracted from one large-scale passive data source. However, due to improved quality and possibly a greater level of detail, more applications may be possible after fusing several data sources.

Another limitation of using travel patterns extracted from large-scale passive data directly is that we can use the data to understand historical and current patterns but not forecast the effect of changes to the traffic system. Forecasts and comparisons of the effects of different traffic system changes require using traffic models (see Chapter 2.4).

Traditionally, traffic models have been based on census data, traffic counts and travel surveys. We could use travel patterns extracted from large-scale passive data to improve the quality and detail of traffic models. Two approaches may be used to improve traffic models using new data sources. The first approach is to use extracted travel patterns as input to traffic models. For example, we could use travel demand estimated from cellular network data as an input to a mode choice or route choice model. Some authors also propose to design new traffic models designed to take large-scale passive data as input (Bwambale et al., 2017). The second approach of integrating the extracted travel patterns with traffic models is to use the data to calibrate traffic models instead of using the travel patterns directly as input. The calibration of travel behaviour models, such as models for mode and route choice, often suffers from the limited sample size of travel surveys. Cellular network data and smart card data can be used to obtain large samples of observations, which we can use to fit the parameters of a model more exactly (Jánošíková et al., 2014). The large sample size of observations could also allow making travel behaviour models more detailed and include more parameters such as dynamic events and the influence of weather as in Zannat and Choudhury (2019). In particular, for agent-based traffic models, which in detail model individual travel patterns, large-scale observations of real travel patterns make it possible to calibrate and validate the models in detail. Potentially, this could lead to much more detailed and accurate traffic models than the traditional four-step traffic models used in practice today.

Chapter 5

Summary of research

This chapter summarises the research of this thesis. It introduces the research questions that this thesis aims to answer and which have been formulated based on the identified research gaps. The general research setup used to conduct the research is introduced. Finally, the included papers are summarised and condensed into a number of main contributions.

5.1 Research Gaps

A considerable number of studies proposing methods of extracting travel patterns from large-scale passive data have already been conducted, as discussed in Chapter 4. However, several relevant research gaps still prevent or limit the use of large-scale passive data for practical traffic planning applications.

First, there are still some relevant open problems where only limited research effort has been made. When it comes to cellular network data, for example, much focus has been on the extraction of trips and travel demand estimation. However, additional metadata needs to be inferred for many applications, such as the route and travel mode used. Existing studies on mode classification of trips extracted from cellular network data are often using simple rule-based algorithms leaving room for new, more advanced methods (Huang et al., 2019). Regarding smart card data, several studies have proposed methods to study travel behaviour. Most focus, however, has been on short-term disruptions while large system changes and long-term planned disrup-

tions, for example, due to construction works, are understudied.

Second, many studies using large-scale passive data to extract travel patterns provide only very limited validation. Often the validation is only performed on a very aggregated level, leaving open questions concerning the quality of the individual travel patterns extracted and the limitations of the suggested methods (Chen et al., 2016). To demonstrate that the quality of extracted travel patterns is good enough to be used in practical applications, validation methods need to be advanced and also show that the results are correct on an individual level.

Third, there is a lack of studies using real-world, large-scale passive datasets that demonstrate if the proposed methods not only work in theory, using simulated data or in a small selected case study. Before a method can be applied in practice, it needs to be demonstrated that it can handle large amounts of data of varying quality and resolution. Many earlier studies using cellular network data are using sparse CDR datasets. Today, cellular network data of higher resolution is available, which opens new possibilities but at the same time calls for new methods adapted for the data that is available today.

5.2 Research Questions

Based on the existing research gaps, the following research questions have been formulated:

RQ1 How can large-scale passive data such as cellular network data and smart card data be processed in order to extract travel patterns for traffic planning applications?

RQ2 How can the quality of travel patterns extracted from cellular network data be evaluated?

RQ3 What potential and limitations of using large-scale passive data for travel pattern analysis can be identified from applying methods of processing the data to real-world datasets?

5.3 Research Setup

The research conducted within this thesis involves extensive data processing. A research setup for data-driven analysis has been developed

within the thesis work to enable this data processing. The following requirements have been identified that the research setup needs to fulfil. First, the setup needs to allow executing queries needed to process the data, including spatial queries and GIS operations. It should also allow flexible analysis and data visualisation. Second, the setup needs to allow for the processing of large-scale datasets in a reasonable time. Third, the setup needs to preserve the privacy of individuals. Fourth, the results should be reproducible, and it needs to be transparent how the results have been produced.

The used setup differs slightly between the papers in the thesis. The setup used for Paper V uses the Spark framework[1]. The setup for all other papers is based on a library[2] made by the author of this thesis combined with a PostgreSQL[3] database. Both are based on the concept of MapReduce to parallelise the data processing to increase performance and use all available CPU power when processing large-scale datasets (see Chapter 4.1). The database allows running efficient queries by indexing the data. Spatial queries are performed using the PostGIS extension[4]. The methods and algorithms part of the papers in this thesis have been implemented in Python[5] and use database queries to fetch data and write back results.

To preserve individual privacy when processing large-scale passive data, the approach of limited release (see Chapter 4.6) is used for Paper II. For the remaining papers I, III, IV and V a remote access setup is used. For Paper V, a single remote access server has been used to process the data. For Paper I, III and IV the setup illustrated in Figure 5.1 has been used which is based on the idea to "bring the code to the data". That means that the computation server used to process large-scale cellular network data is controlled by the operator only and cannot be accessed directly. The code for data processing is submitted to the server and executed. Only the aggregated results that do not allow to draw any conclusions about individuals are exported.

Without direct access to the server, debugging and verification of the code is difficult. Therefore, a second server is used for development purposes. Here, the same computational data processing code can be run on a small test dataset collected by a few users who have

[1] https://spark.apache.org
[2] https://pypi.org/project/pyparade/
[3] https://www.postgresql.org
[4] https://postgis.net
[5] https://www.python.org

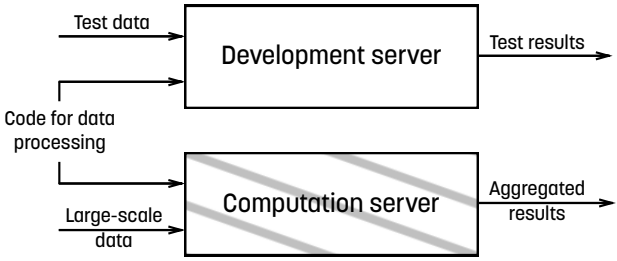


Figure 5.1: The used remote-access setup to calculate aggregated results from cellular network data without direct access to the data. The development server is used for debugging and can be remote-accessed while the computation server is not directly accessible.

given consent to use the data for development purposes. Each run is logged with the exact code revision and parameters used. This makes the computation of the results reproducible.

The R programming language[6] is used in all papers in this thesis to enable the analysis of travel patterns using visualisation, comparison metrics and statistical analysis. Further, QGIS[7] is used to create geospatial visualisations of travel patterns.

5.4 Summary of the Included Papers

The papers in this thesis make contributions to different steps of processing cellular network data for travel pattern analysis (Paper I-IV) and include a method of analysing travel patterns from smart card data (Paper V). All papers consist of a processing and analysis part (see Figure 5.2). This section gives a summary of each of the included papers.

Paper I: *Comparative Analysis of Travel Patterns from Cellular Network Data and an Urban Travel Demand Model*

This paper compares travel demand estimated from cellular network data with an existing travel demand model. A data-driven travel demand estimation method is presented, consisting of a trip extraction step and a travel demand estimation step. We compare two trip

[6] https://www.r-project.org
[7] https://qgis.org/en/site/

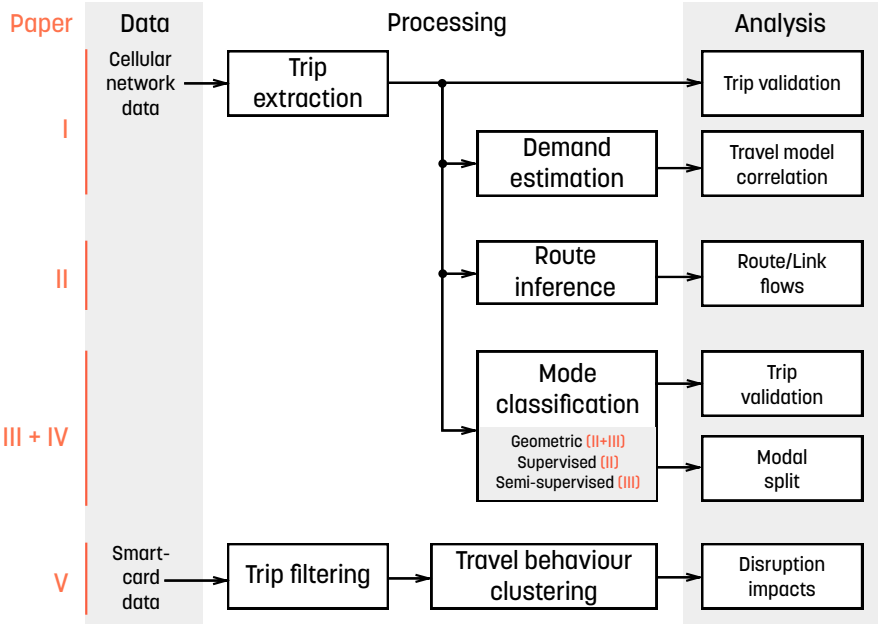


Figure 5.2: Overview of the included papers by data source source, main processing steps and analysis focus.

extraction methods which have been presented earlier in Breyer et al. (2017) regarding their impact on the estimation of travel demand. The bias caused by using different trip extraction methods is further analysed using a small-scale cellular network dataset collected from 20 mobile phones together with GPS tracks collected on the same device is used. We use a large-scale dataset of cellular network data from a Swedish operator for the city of Norrköping to compare the travel demand inferred from cellular network data to the municipality's travel demand model. Additionally, the time profile of the estimated travel demand is compared to a time-sliced variant of the model and public transportation tap-ins.

The results show that the recall is just about 50% for trips which are only 1-2km long while it is 75-80% for trips of more than 5km length. Similarly, the recall also differs by travel mode with more than 80% for public transportation, 74% for car but only 53% for bicycle and walking. The correlation to the municipality's urban travel demand model after aggregating trips into an origin-destination matrix depends on the aggregation level of the comparison. While the correlation is weak ($R^2 < 0.2$) using the original zoning used in the

traditional model with 189 zones, it is good ($R^2 = 0.82$) when aggregating to 24 zones. We find systematic differences in travel demand resulting from two different trip extraction methods. This highlights that the choice of the trip extraction method is crucial for the travel demand estimation.

Paper I is co-authored with David Gundlegård and Clas Rydergren. The author of the thesis is the main author and has contributed with major work, including conceptualisation, method development, analysis and paper writing.

Paper I has been published in

Breyer, N.; Rydergren, C.; Gundlegård, D. (2020). Comparative analysis of travel patterns from cellular network data and an urban travel demand model. *Journal of Advanced Transportation*, 2020

Parts of Paper I have been presented at *Transportforum 2019* in Linköping, Sweden and *Netmob 2019* in Oxford, United Kingdom.

Paper II: *Cellpath Routing and Route Traffic Flow Estimation Based on Cellular Network data*

Paper II focuses on the route inference for trips extracted from cellular network data to estimate traffic flows. We present and compare four algorithms to estimate a route on the road network for a given cellpath of a trip: First, *Direct routing* uses the shortest path between origin and destination of the cellpath. Second, *Waypoint Routing* where the route is forced to go through all Voronoi cells in the cellpath. Third, *Magnetic Routing* uses modified link costs to infer a route that follows the cellpath. Fourth, *Magentic Waypoint Routing* which combines the previous two algorithms.

We compare the four route inference methods regarding the characteristics of the individual routes. To investigate the effect of using the different route estimation methods in a network loading for a larger city, we use a large-scale CDR dataset of Dakar, Senegal. The results show that the choice of the route estimation method has a significant impact on the resulting link flows. In particular, we find that strictly following the cellpath of each trip leads to unrealistic routes.

Paper II is co-authored with David Gundlegård and Clas Rydergren. The author of the thesis is the main author and has contributed

with major work, including conceptualisation, method development, analysis and paper writing.

Paper II has been published in

Breyer, N.; Gundlegård, D.; Rydergren, C. (2018). Cellpath routing and route traffic flow estimation based on cellular network data. *Journal of Urban Technology*, 25(2):85–104. doi: 10.1080/10630732.2017.1386939

Parts of Paper II have been presented at *MobileTartu 2016* in Tartu, Estonia, *Netmob 2017* in Milan, Italy and the *Swedish Transportation Research Conference* in Lund, Sweden.

Paper III: *Travel mode classification of intercity trips using cellular network data*

Some applications in transport planning require knowing the travel mode of trips. Classification by travel mode is needed to use cellular network data in these applications. Paper III involves a comparison of three geometry-based mode classification methods and three supervised learning methods to classify trips previously extracted from cellular network data in intercity origin-destination pairs as either made by road or rail.

To compare the different classification methods, we use a labelled dataset of 255 trips in two inter-city OD-pairs in Sweden to train the supervised classification methods and evaluate the classification performance. For an OD-pair where the road and train routes are not separated by more than four kilometres, the geometry-based methods classify 4.5%–7.1% of the trips wrong, while two of the supervised learning methods can classify all trips correctly. Using a large-scale dataset of 29037 trips, we find that separation between classes is less clear than in the labelled dataset. We also show how the choice of classification method impacts the aggregated modal split estimate.

Paper III is co-authored with David Gundlegård and Clas Rydergren. The author of the thesis is the main author and has contributed with major work, including conceptualisation, method development, analysis and paper writing.

Paper III has been published in

Breyer, N.; Gundlegård, D.; Rydergren, C. (2021). Travel mode classification of intercity trips using cellular network data. *Transportation Research Procedia*, 52:211 – 218. doi: 10.1016/j.trpro.2021.01.024. 23rd EURO Working Group on Transportation Meeting, EWGT 2020, 16-18 September 2020, Paphos, Cyprus

Parts of Paper III have been presented at *Mobile Tartu 2020* in Tartu, Estonia (virtual conference), the *23rd Euro Working Group on Transportation* in Paphos, Cyprus (virtual conference) and the *Swedish Transportation Research Conference* in Karlstad, Sweden (virtual conference).

Paper IV: *Semi-supervised mode classification of inter-city trips from cellular network data*

As Paper III, also Paper IV is focusing on the problem of classifying trips extracted from cellular network data by travel mode. While the supervised methods presented in Paper III require a set of manually labelled trips for training, Paper IV aims to classify without any labelled training data as this is usually not available in practice. The paper presents first a geometric classification method, which in contrast to Paper II also works when a user takes another route then the fasted route alternative when travelling by road.

The main contributions of Paper IV are three methods of classification through semi-supervised labelling. The methods label first a few trips using geometric classification by comparing to route alternatives. The semi-supervised labelling approaches use then underlying structures in a large set of unlabelled trips to label the remaining trips by using a standard supervised classification method (geometric-labelling), iterative semi-supervised training (self-labelling) or by transferring information between OD-pairs (continuity-labelling).

We apply the three semi-supervised classification methods on a dataset of 9545 unlabelled trips in two inter-city origin-destination pairs. We find that the methods can identify structures in the cells used during trips in the unlabelled data corresponding to the available route alternatives. Finally, we validate the classification methods using a dataset of 255 manually labelled trips in the two origin-destination pairs. While geometric classification misclassified 4.2% and 5.6% of the trips in the two origin-destination pairs, all trips can be classified correctly using semi-supervised classification.

Paper IV is co-authored with David Gundlegård and Clas Rydergren. The author of the thesis is the main author and has contributed with major work, including conceptualisation, method development, analysis and paper writing.

Paper IV is a working paper.

Paper V: *Analysing the impacts of long-term service disruptions on passenger travel behaviour: A smart card analysis from the Greater Copenhagen area*

Paper V contains an analysis of travel patterns using smart card data in order to understand the impacts of long-term disruption in public transportation. While previous studies have focused on short-term disruptions or only used aggregated data, we propose a new method based on smart card data for analysing the impacts of long-term planned disruptions on individual passenger travel behaviour. We use k-means clustering to group passengers based on their travel behaviour before and after the closure. Thus, we can observe how different passenger groups changed travel behaviour after the disruption and compare these observations to a reference line without disruption to account for general trends. Using hierarchical clustering of daily travel patterns, we can analyse certain passenger groups' reactions to the disruption.

We apply the method on a 3-month closure of a rail line in the Greater Copenhagen area. The results suggest that, in particular, passengers with everyday commuting behaviour have decreased after the disruption. As this group stands for a large portion of the total trips, this contributes to the lower ridership on the disrupted line after the disruption compared to an unaffected reference line. The proposed methodology enables the analysis of the impact of disruptions on diverse passengers segments which is useful for public transport agencies when planning long-term maintenance projects.

Paper V is co-authored with Morten Eltved, Jesper Bláfoss Ingvardson and Otto Anker Nielsen. The author of the thesis has contributed as co-author and with major work regarding the conceptualisation and implementation of the analysis of impacts on different passenger groups using clustering by travel behaviour (in particular Sections 3.1-3.3, 4.2 and 5.1-5.3).

Paper V is a working paper. Parts of Paper V have been presented at *TransitData 2020* in Toronto, Canada (virtual conference).

5.5 Other Related Work by the Author

In addition to the five papers included in this thesis, the author has also contributed to two related papers not included in this thesis. The first of these shows how CDR data can be used to estimate travel demand. It was excluded due to the overlap with Paper I and as the other co-authors have done major work in the paper. The paper has been published in

Gundlegård, D.; Rydergren, C.; Breyer, N.; Rajna, B. (2016). Travel demand estimation and network assignment based on cellular network data. *Computer Communications*, 95:29 – 42. doi: 10.1016/j.comcom.2016.04.015

The second paper focuses on the trip extraction from cellular network data and how the resulting trips can be validated. The paper is not included in this thesis due to the overlap with Paper I. The paper has been published in

Breyer, N.; Gundlegård, D.; Rydergren, C.; Bäckman, J. (2017). Trip extraction for traffic analysis using cellular network data. In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 321–326. doi: 10.1109/MTITS.2017.8005688

5.6 Main Contributions

The main contributions of this thesis can be summarised as follows:

1 Methods for extracting trips and estimate travel demand from noisy cellular network data of low resolution in space and time (Paper I)

2 Methods for inferring routes of trips extracted from cellular network data (Paper II)

3 Methods for classifying trips extracted from cellular network data by travel mode by means of rule-based and machine learning methods and using structures in large-scale data (Paper III and IV)

4 A method for analysing travel behaviour changes using clustering of smart card data (Paper V)

5 Approaches for validating extracted travel patterns on individual level (Paper I, III, IV) and aggregated level (Paper I)

6 Approaches for evaluating the impact of using different methods for extracting travel patterns (Paper I, II, III, IV)

7 Identification of potential and limitations in practical applications of large-scale passive data for travel pattern analysis using real-world large-scale datasets (Chapters 3, 4, 6 and Papers I-V)

While contributions 1-3, 5 and 6 are covered in this thesis for cellular network data, contribution 4 is covered using smart card data. Contribution 7 is related to the use of both data sources. The contributions relate to the research questions introduced in Chapter 5.2) as follows. Contributions 1-4 are related to research question RQ1 by proposing different methods to process large-scale passive data and correspond to the processing part in Figure 5.2. Contributions 5 and 6 are related to RQ2 and correspond to the different methods of analysing and evaluating the results in Figure 5.2. Finally, contribution 7 relates to RQ3 and summarises the knowledge on the potential and limitations of different large-scale passive data sources as a result of using different real-world data sets in the papers of this thesis.

Chapter 6

Conclusions and Future Work

Large-scale passive data sources allow obtaining large samples of observations of travel patterns. Cellular network data collected by a cellular network operator can provide a unique overview of travel patterns with all travel modes for a significant share of the population. Smart card data covers travel patterns in public transportation with a great level of detail. Before these data sources can help traffic planners, extensive data processing is needed to extract useful travel patterns. This thesis extended the knowledge about suitable data processing methods.

The findings in this thesis include that it is important to take the characteristics of the data into account when designing a data processing method to extract travel patterns from large-scale passive data. Paper II illustrated that route inference from CDRs by strictly following the cellpath is problematic due to the noise and low resolution of the data. Both rule-based and machine learning methods can be used to extract travel patterns. Rule-based methods are often easier to understand and are appropriate for problems where the patterns to be extracted can be described clearly. Paper I shows that a simple rule-based stop detection algorithm can be used to extract longer trips from cellular network data reliably. However, rule-based methods are problematic when the patterns to be detected are complex or not exactly known beforehand. In these cases, machine learning methods can be used. Unsupervised learning can be used to find patterns in

the data without prior specification. Paper V showed how clustering of smart card data could be used to group public transportation users by travel behaviour to understand the effects of a disruption. Paper III showed that for the problem of travel mode classification of trips extracted from cellular network data supervised classification can outperform rule-based geometric methods. Supervised learning requires training data to learn patterns. When no or little training data is available, using semi-supervised learning is a promising approach as demonstrated in Paper IV.

Each data processing method needs to be evaluated to understand its performance. Three approaches for evaluation are proposed in the thesis. First, the validation on individual level using data collected in experiments, such that the correct travel patterns are known (Papers I, III and IV). Second, the comparison with other independent data sources, such as an existing traffic model, to understand whether the extracted travel patterns match the expectations of travel patterns on an aggregated level (Paper I). Third, the comparison of methods to analyse the impact of using different methods to process the same data (Papers I-IV). These evaluation approaches complement each other. Since datasets collected in experiments are rarely representative of the population, it may not be enough to conclude on the quality of the extracted travel patterns on an aggregated level. The comparison with another independent data source, on the other hand, does not allow us to conclude whether the individual travel patterns detected are correct. Using the comparison of methods, Papers I-IV show that the choice of the data processing method can have large effects on the results.

In the studies of this thesis, real-world, large-scale passive datasets have been used to demonstrate how the extraction of travel patterns could work under realistic circumstances. This has exposed two main limitations of using large-scale passive data for travel pattern analysis. First, there are limitations due to the particular large-scale passive data source used. Cellular network data for example, is limited by its resolution in time and space, which makes it impossible to capture very short trips reliably in areas where the cell density is low (Paper I). Second, even though large-scale passive data sources cover large portions of the population, there may be sample bias. No socio-economic data is usually available linked to individuals in the data, making it challenging to control this bias. At the same time, the studies of this thesis have shown the potential of using large-scale

passive data. Due to the large sample size, the data allows understanding travel patterns based on observations instead of relying on traffic models' underlying assumptions and parameters. The data can be updated much easier than, for example, a travel survey. This allows to quickly understand travel behaviour changes that could not be detected otherwise.

Despite the active research in the area of using large-scale passive data for travel pattern analysis still, some open problems remain. Regarding data processing methods, the problem of classifying activities from cellular network data has, for example, not been studied extensively. The problem of scaling travel demand to the whole population may also need more attention. While large-scale passive data theoretically allows understanding travel patterns in real-time, few methods have been designed to take advantage of this. Better metrics are needed to establish ways of quantifying the similarity of travel patterns from different sources. Simple correlation metrics often cannot describe the similarity of travel patterns sufficiently. Practical traffic planning applications could also benefit from the data fusion of several data sources and the integration of travel patterns extracted from large-scale passive data with traffic models as input or for calibration.

Bibliography

Aggarwal, A.; Guibas, L. J.; Saxe, J.; Shor, P. W. (1989). A linear-time algorithm for computing the voronoi diagram of a convex polygon. *Discrete & Computational Geometry*, 4(1):591–604.

Ahas, R.; Aasa, A.; Mark, Ü.; Pae, T.; Kull, A. (2007). Seasonal tourism spaces in estonia: Case study with mobile positioning data. *Tourism management*, 28(3):898–910.

Alessandretti, L.; Aslak, U.; Lehmann, S. (2020). The scales of human mobility. *Nature*, 587(7834):402–407.

Alexander, L.; Jiang, S.; Murga, M.; González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58, Part B:240 – 250. doi: 10.1016/j.trc.2015.02.018. Big Data in Transportation and Traffic Engineering.

Algizawy, E.; Ogawa, T.; El-Mahdy, A. (2017). Real-time large-scale map matching using mobile phone data. *ACM Trans. Knowl. Discov. Data*, 11(4):52:1–52:38. doi: 10.1145/3046945.

Alsger, A.; Tavassoli, A.; Mesbah, M.; Ferreira, L.; Hickman, M. (2018). Public transport trip purpose inference using smart card fare data. *Transportation Research Part C: Emerging Technologies*, 87:123–137. doi: 10.1016/j.trc.2017.12.016.

Alsger, A. A.; Mesbah, M.; Ferreira, L.; Safi, H. (2015). Use of smart card fare data to estimate public transport origin-destination matrix. *Transportation Research Record*, 2535(1):88–96. doi: 10.3141/2535-10.

Anda, C.; Erath, A.; Fourie, P. J. (2017a). Transport modelling in the age of big data. *International Journal of Urban Sciences*, 21(sup1):19–42. doi: 10.1080/12265934.2017.1281150.

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

Anderson, J. A. (1995). *An introduction to neural networks*. MIT press.

Bachir, D.; Khodabandelou, G.; Gauthier, V.; El, M. (2018). Combining bayesian inference and clustering for transport mode detection from sparse and noisy geolocation data. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2018*.

Bachir, D.; Khodabandelou, G.; Gauthier, V.; El Yacoubi, M.; Puchinger, J. (2019). Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies*, 101:254–275.

Baert, A.-E.; Seme, D. (2004). Voronoi mobile cellular networks: topological properties. In *Parallel and Distributed Computing, 2004. Third International Symposium on/Algorithms, Models and Tools for Parallel Computing on Heterogeneous Networks, 2004. Third International Workshop on*, pages 29–35. doi: 10.1109/ISPDC.2004.58.

Balakrishna, R.; Antoniou, C.; Ben-Akiva, M.; Koutsopoulos, H.; Wen, Y. (2015). Calibration of microscopic traffic simulation models: Methods and application. *Transportation Research Record: Journal of the Transportation Research Board*.

Balmer, M.; Rieser, M.; Meister, K.; Charypar, D.; Lefebvre, N.; Nagel, K.; Axhausen, K. (2009). Matsim-t: Architecture and simulation times. *Multi-agent systems for traffic and transportation engineering*, pages 57–78.

Barbosa, H.; Barthelemy, M.; Ghoshal, G.; James, C. R.; Lenormand, M.; Louail, T.; Menezes, R.; Ramasco, J. J.; Simini, F.; Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, 734:1 – 74. doi: 10.1016/j.physrep.2018.01.001. Human mobility: Models and applications.

Batran, M.; Mejia, M. G.; Sekimoto, Y.; Shibasaki, R. (2018). Inference of human spatiotemporal mobility in greater maputo by mobile phone big data mining. In *Proceedings of the Tenth International Workshop on Agents in Traffic and Transportation*.

Becker, R. A.; Caceres, R.; Hanson, K.; Loh, J. M.; Urbanek, S.; Varshavsky, A.; Volinsky, C. (2011a). Route classification using cellular handoff patterns. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 123–132. Association for Computing Machinery, New York, NY, USA. ISBN 9781450306300. doi: 10.1145/2030112.2030130.

Becker, R. A.; Caceres, R.; Hanson, K.; Loh, J. M.; Urbanek, S.; Varshavsky, A.; Volinsky, C. (2011b). A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):0018–26.

Bernhardt, K. (2007). Agent-based modeling in transportation. *Artificial Intelligence in Transportation*, 72.

Bhaskaran, H.; Raychaudhuri, D.; Verma, S. (2003). Capacity analysis of a cellular data system with 3g/wlan interworking. In *2003 IEEE 58th Vehicular Technology Conference. VTC 2003-Fall (IEEE Cat. No.03CH37484)*, volume 3, pages 1817–1821 Vol.3. doi: 10.1109/VETECF.2003.1285339.

Brakewood, C.; Kocur, G. (2011). Modeling transit rider preferences for contactless bank cards as fare media: transport for london and the chicago, illinois, transit authority. *Transportation research record*, 2216(1):100–107.

Breyer, N.; Gundlegård, D.; Rydergren, C. (2018). Cellpath routing and route traffic flow estimation based on cellular network data. *Journal of Urban Technology*, 25(2):85–104. doi: 10.1080/10630732.2017.1386939.

Breyer, N.; Gundlegård, D.; Rydergren, C. (2021). Travel mode classification of intercity trips using cellular network data. *Transportation Research Procedia*, 52:211 – 218. doi: 10.1016/j.trpro.2021.01.024. 23rd EURO Working Group on Transportation Meeting, EWGT 2020, 16-18 September 2020, Paphos, Cyprus.

Breyer, N.; Gundlegård, D.; Rydergren, C.; Bäckman, J. (2017). Trip extraction for traffic analysis using cellular network data. In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 321–326. doi: 10.1109/MTITS.2017.8005688.

Breyer, N.; Rydergren, C.; Gundlegård, D. (2020). Comparative analysis of travel patterns from cellular network data and an urban travel demand model. *Journal of Advanced Transportation*, 2020.

Briand, A.-S.; CÃŽme, E.; TrÃ©panier, M.; Oukhellou, L. (2017). Analyzing year-to-year changes in public transport passenger behaviour using smart card data. *Transportation Research Part C: Emerging Technologies*, 79:274 – 289. doi: 10.1016/j.trc.2017.03.021. Cluster analysis of passengers based on temporal patterns.

Bwambale, A.; Choudhury, C. F.; Hess, S. (2017). Modelling trip generation using mobile phone data: A latent demographics approach. *Journal of Transport Geography*. doi: 10.1016/j.jtrangeo.2017.08.020.

Caceres, N.; Wideberg, J. P.; Benitez, F. G. (2008). Review of traffic data estimations extracted from cellular networks. *IET Intelligent Transport Systems*, 2(3):179–192. Copyright - Copyright The Institution of Engineering & Technology Sep 2008; Document feature - Diagrams; Graphs; Illustrations; Tables; ; Last updated - 2014-12-18.

Calabrese, F.; Di Lorenzo, G.; Liu, L.; Ratti, C. (2011). Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, 10(4):36.

Calabrese, F.; Diao, M.; Lorenzo, G. D.; Jr., J. F.; Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26:301 – 313. doi: 10.1016/j.trc.2012.09.009.

Calabrese, F.; Ferrari, L.; Blondel, V. D. (2014). Urban sensing using mobile phone network data: A survey of research. *ACM Comput. Surv.*, 47(2):25:1–25:20. doi: 10.1145/2655691.

Calabrese, F.; Pereira, F. C.; Di Lorenzo, G.; Liu, L.; Ratti, C. (2010). The geography of taste: Analyzing cell-phone mobility and social events. In *Proceedings of the 8th International Conference on Pervasive Computing*, Pervasive'10, pages 22–37. Springer-Verlag, Berlin, Heidelberg. ISBN 3-642-12653-7, 978-3-642-12653-6. doi: 10.1007/978-3-642-12654-3_2.

Cao, Q.; Ren, G.; Li, D.; Ma, J.; Li, H. (2020). Semi-supervised route choice modeling with sparse automatic vehicle identification data. *Transportation Research Part C: Emerging Technologies*, 121:102857.

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

Chin, K.; Huang, H.; Horn, C.; Kasanicky, I.; Weibel, R. (2019). Inferring fine-grained transport modes from mobile phone cellular signaling data. *Computers, Environment and Urban Systems*, 77:101348. doi: 10.1016/j.compenvurbsys.2019.101348.

Cho, E.; Myers, S. A.; Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM.

Chu, K. K. A.; Chapleau, R. (2008). Enriching archived smart card transaction data for transit demand modeling. *Transportation Research Record*, 2063(1):63–72. doi: 10.3141/2063-08.

Dash, M.; Koo, K. K.; Holleczek, T.; Yap, G. E.; Krishnaswamy, S. P.; Shi-Nash, A. (2015). From mobile phone data to transport network – gaining insight about human mobility. In *2015 16th IEEE International Conference on Mobile Data Management*, volume 1, pages 243–250. doi: 10.1109/MDM.2015.74.

de Montjoye, Y.-A.; Gambs, S.; Blondel, V.; Canright, G.; de Cordes, N.; Deletaille, S.; Engø-Monsen, K.; Garcia-Herranz, M.; Kendall, J.; Kerry, C.; Krings, G.; Letouzé, E.; Luengo-Oroz, M.; Oliver, N.; Rocher, L.; Rutherford, A.; Smoreda, Z.; Steele, J.; Wetter, E.; Pentland, A. S.; Bengtsson, L. (2018). On the privacy-conscientious use of mobile phone data. *Scientific Data*, 5:180286 EP –.

Deschaintres, E.; Morency, C.; TrÃ©panier, M. (2019). Analyzing transit user behavior with 51 weeks of smart card data. *Transportation Research Record: Journal of the Transportation Research Board*, 2673:33–45. doi: 10.1177/0361198119834917.

Dong, H.; Wu, M.; Ding, X.; Chu, L.; Jia, L.; Qin, Y.; Zhou, X. (2015). Traffic zone division based on big data from mobile phone base stations. *Transportation Research Part C: Emerging Technologies*, 58:278 – 291. doi: 10.1016/j.trc.2015.06.007. Big Data in Transportation and Traffic Engineering.

Egu, O.; Bonnel, P. (2020). Investigating day-to-day variability of transit usage on a multimonth scale with smart card data. a case study in lyon. *Travel Behaviour and Society*, 19:112–123.

Erlander, S.; Stewart, N. F. (1990). *The gravity model in transportation analysis: theory and extensions*, volume 3. Vsp.

Fillekes, M. (2014). *Reconstructing Trajectories from Sparse Call Detail Records*. Master's thesis, University of Tartu.

Gonzalez, M. C.; Hidalgo, C. A.; Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Gundlegård, D. (2018). *Transport Analytics Based on Cellular Network Signalling Data*. Ph.D. thesis, Linköping University, Communications and Transport Systems, Faculty of Science & Engineering. doi: 10.3384/diss.diva-152237.

Gundlegård, D.; Rydergren, C.; Breyer, N.; Rajna, B. (2016). Travel demand estimation and network assignment based on cellular network data. *Computer Communications*, 95:29 – 42. doi: 10.1016/j.comcom.2016.04.015.

Haghani, A.; Hamedi, M.; Sadabadi, K. F.; Young, S.; Tarnoff, P. (2010). Data collection of freeway travel time ground truth with bluetooth sensors. *Transportation Research Record*, 2160(1):60–68.

Han, G.; Sohn, K. (2016). Activity imputation for trip-chains elicited from smart-card data using a continuous hidden markov model. *Transportation Research Part B: Methodological*, 83:121–135.

Hartmanis, J.; Stearns, R. E. (1965). On the computational complexity of algorithms. *Transactions of the American Mathematical Society*, 117:285–306.

Hasan, S.; Zhan, X.; Ukkusuri, S. V. (2013). Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, UrbComp '13. Association for Computing Machinery, New York, NY, USA. ISBN 9781450323314. doi: 10.1145/2505821.2505823.

He, L.; Trépanier, M. (2015). Estimating the destination of unlinked trips in transit smart card fare data. *Transportation Research Record*, 2535(1):97–104. doi: 10.3141/2535-11.

Hofleitner, A.; Herring, R.; Abbeel, P.; Bayen, A. (2012). Learning the dynamics of arterial traffic from probe data using a dynamic bayesian network. *IEEE Transactions on Intelligent Transportation Systems*, 13(4):1679–1693. doi: 10.1109/TITS.2012.2200474.

Huang, H.; Cheng, Y.; Weibel, R. (2019). Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*.

Huang, Z.; Huang, Z.; Zheng, P.; Xu, W. (2018). Calibration of c-logit-based sue route choice model using mobile phone data. *Information*, 9(5). doi: 10.3390/info9050115.

Isaacman, S.; Becker, R.; Cáceres, R.; Kobourov, S.; Martonosi, M.; Rowland, J.; Varshavsky, A. (2011). Identifying important places in people's lives from cellular network data. In K. Lyons; J. Hightower; E. Huang, editors, *Pervasive Computing*, volume 6696 of *Lecture Notes in Computer Science*, pages 133–151. Springer Berlin Heidelberg. ISBN 978-3-642-21725-8. doi: 10.1007/978-3-642-21726-5_9.

Jagadeesh, G. R.; Srikanthan, T. (2017). Online map-matching of noisy and sparse location data with hidden markov and route choice models. *IEEE Transactions on Intelligent Transportation Systems*, 18(9):2423–2434. doi: 10.1109/TITS.2017.2647967.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jánošíková, L.; Slavík, J.; Koháni, M. (2014). Estimation of a route choice model for urban public transport using smart card data. *Transportation planning and technology*, 37(7):638–648.

Jaume, B.; Montero Mercadé, L.; Bullejos, M.; Serch, O.; Carmona, C. (2012). A kalman filter approach for the estimation of time dependent od matrices exploiting bluetooth traffic data collection. In *TRB 91st Annual Meeting Compendium of Papers DVD*, pages 1–16.

Jung, J.; Sohn, K. (2017). Deep-learning architecture to forecast destinations of bus passengers from entry-only smart-card data. *IET Intelligent Transport Systems*, 11(6):334–339.

Kaisler, S.; Armour, F.; Espinosa, J. A.; Money, W. (2013). Big data: Issues and challenges moving forward. In *2013 46th Hawaii international conference on system sciences*, pages 995–1004. IEEE.

Kalatian, A.; Shafahi, Y. (2016). Travel mode detection exploiting cellular network data. *MATEC Web Conf.*, 81:03008. doi: 10.1051/matecconf/20168103008.

Kanasugi, H.; Sekimoto, Y.; Kurokawa, M.; Watanabe, T.; Muramatsu, S.; Shibasaki, R. (2013). Spatiotemporal route estimation consistent with human mobility using cellular network data. In *2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)*, pages 267–272. doi: 10.1109/PerComW.2013.6529493.

Kazagli, E.; Koutsopoulos, H. N. (2013). Estimation of arterial travel time from automatic number plate recognition data. *Transportation research record*, 2391(1):22–31.

Leontiadis, I.; Lima, A.; Kwak, H.; Stanojevic, R.; Wetherall, D.; Papagiannaki, K. (2014). From cells to streets: Estimating mobile paths with cellular-side data. In *Proceedings of the 10th ACM International on Conference on emerging Networking Experiments and Technologies*, pages 121–132. ACM, ACM.

Liero, H.; Zwanzig, S. (2016). *Introduction to the Theory of Statistical Inference*. Chapman & Hall/CRC Texts in Statistical Science. CRC Press. ISBN 9781466503205.

Liu, F.; Janssens, D.; Cui, J.; Wang, Y.; Wets, G.; Cools, M. (2014). Building a validation measure for activity-based transportation models based on mobile phone data. *Expert Systems with Applications*, 41(14):6174 – 6189. doi: 10.1016/j.eswa.2014.03.054.

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

Manley, E.; Zhong, C.; Batty, M. (2018). Spatiotemporal variation in travel regularity through transit user profiling. *Transportation*, 45(3):703–732.

Ming-Heng, W.; Schrock, S.; Broek, N.; Mulinazzi, T. (2013). Estimating dynamic origin-destination data and travel demand using cell phone network data. *International Journal of Intelligent Transportation Systems Research*, 11(2):76 – 86.

Mojica, C. H. (2008). Examining changes in transit passenger travel behavior through a smart card activity analysis (master thesis), massachusetts institute of technology.

Morency, C.; Trepanier, M.; Agard, B. (2007). Measuring transit performance using smart card data. In *World Conference on Transport Research, San Francisco, USA*.

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

Östh, J.; Shuttleworth, I.; Niedomysl, T. (2018). Spatial and temporal patterns of economic segregation in sweden's metropolitan areas: A mobility approach. *Environment and Planning A: Economy and Space*, 50(4):809–825.

Patriksson, M. (2015). *The Traffic Assignment Problem: Models and Methods*. Dover Publications. ISBN 9780486802275.

Pavlo, A.; Paulson, E.; Rasin, A.; Abadi, D. J.; DeWitt, D. J.; Madden, S.; Stonebraker, M. (2009). A comparison of approaches to large-scale data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, SIGMOD '09,

pages 165–178. Association for Computing Machinery, New York, NY, USA. ISBN 9781605585512. doi: 10.1145/1559845.1559865.

Pelletier, M.-P.; TrÃ©panier, M.; Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies*, 19:557 – 568. doi: 10.1016/j.trc.2010.12.003.

Phithakkitnukoon, S.; Sukhvibul, T.; Demissie, M.; Smoreda, Z.; Natwichai, J.; Bento, C. (2017). Inferring social influence in transport mode choice using mobile phone data. *EPJ Data Science*, 6(1):11.

Pollard, T.; Taylor, N.; van Vuren, T.; MacDonald, M. (2013). Comparing the quality of od matrices in time and between data sources. In *Proceedings of the European Transport Conference*.

Prelipcean, A. C.; Gidófalvi, G.; Susilo, Y. O. (2015). Comparative framework for activity-travel diary collection systems. In *2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 251–258.

Qu, Y.; Gong, H.; Wang, P. (2015). Transportation mode split with mobile phone data. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 285–289. doi: 10.1109/ITSC.2015.56.

Rabiner, L.; Juang, B. (1986). An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16. doi: 10.1109/MASSP.1986.1165342.

Saifullah, Y.; Zheng, H.; Maheshwari, S. (2012). Handover or location update for optimization for relay stations in a wireless network. US Patent 8,140,077.

Sari Aslam, N.; Cheng, T.; Cheshire, J. (2019). A high-precision heuristic model to detect home and work locations from smart card data. *Geo-spatial Information Science*, 22(1):1–11.

Schlaich, J. (2010). Analyzing route choice behavior with mobile phone trajectories. *TRANSPORTATION RESEARCH RECORD*, 2(2157):78–85.

Schulz, A.; Nobis, C.; Eggs, J.; Bäumer, M. (2016). German national travel survey 'mid 2016 – mobility in germany': New challenges – new approaches. In *European Transport Conference 2016*, AET Papers Repository.

Song, X.; Ouyang, Y.; Du, B.; Wang, J.; Xiong, Z. (2017). Recovering individual's commute routes based on mobile phone data. *Mobile Information Systems*, 2017.

Stopher, P.; FitzGerald, C.; Xu, M. (2007). Assessing the accuracy of the sydney household travel survey with gps. *Transportation*, 34(6):723–741. doi: 10.1007/s11116-007-9126-8.

Tettamanti, T.; Demeter, H.; Varga, I. (2012). Route choice estimation based on cellular signaling data. *Acta Polytechnica Hungarica*, 9(4):207–220.

Toch, E.; Lerner, B.; Ben-Zion, E.; Ben-Gal, I. (2018). Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems*. doi: 10.1007/s10115-018-1186-x.

Toole, J. L.; Colak, S.; Sturt, B.; Alexander, L. P.; Evsukoff, A.; González, M. C. (2015). The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 58, Part B:162 – 177. doi: 10.1016/j.trc.2015.04.022. Big Data in Transportation and Traffic Engineering.

Treiber, M.; Kesting, A. (2013). Traffic flow dynamics. *Traffic Flow Dynamics: Data, Models and Simulation, Springer-Verlag Berlin Heidelberg*.

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

Tsanakas, N.; Ekström, J.; Gundlegård, D.; Olstam, J.; Rydergren, C. (2021). Data-driven network loading. *Transportmetrica B: Transport Dynamics*, 9(1):237–265.

van Engelen, J. E.; Hoos, H. H. (2019). A survey on semi-supervised learning. *Machine Learning*, 109(2):373–440. doi: 10.1007/s10994-019-05855-6.

Vanhoof, M.; Reis, F.; Ploetz, T.; Smoreda, Z. (2018). Assessing the quality of home detection from mobile phone data for official statistics. *Journal of Official Statistics*, 34(4):935 – 960.

Viallard, A.; Trépanier, M.; Morency, C. (2019). Assessing the evolution of transit user behavior from smart card data. *Transportation Research Record*, 2673(4):184–194.

Wall, M. E.; Rechtsteiner, A.; Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer.

Wang, H.; Calabrese, F.; Di Lorenzo, G.; Ratti, C. (2010). Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, pages 318–323. doi: 10.1109/ITSC.2010.5625188.

Wang, Z.; He, S. Y.; Leung, Y. (2018). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, 11:141 – 155. doi: 10.1016/j.tbs.2017.02.005.

Wen, C.-H.; Koppelman, F. S. (2001). The generalized nested logit model. *Transportation Research Part B: Methodological*, 35(7):627 – 641. doi: 10.1016/S0191-2615(00)00045-X.

Widhalm, P.; Yang, Y.; Ulm, M.; Athavale, S.; González, M. C. (2015). Discovering urban activity patterns in cell phone data. *Transportation*, 42(4):597–623. doi: 10.1007/s11116-015-9598-x.

Wismans, L.; Friso, K.; Rijsdijk, J.; de Graaf, S.; Keij, J. (2018). Improving a priori demand estimates transport models using mobile phone data: A rotterdam-region case. *Journal of Urban Technology*, 25(2):63–83. doi: 10.1080/10630732.2018.1442075.

Wu, C.; Thai, J.; Yadlowsky, S.; Pozdnoukhov, A.; Bayen, A. (2015). Cellpath: Fusion of cellular and traffic sensor data for route flow estimation via convex optimization. *Transportation Research Part C: Emerging Technologies*.

Wu, W.; Wang, Y.; Gomes, J. B.; Anh, D. T.; Antonatos, S.; Xue, M.; Yang, P.; Yap, G. E.; Li, X.; Krishnaswamy, S.; Decraene, J.; Shi-Nash, A. (2014). Oscillation resolution for mobile phone

cellular tower data to enable mobility modelling. In *2014 IEEE 15th International Conference on Mobile Data Management*, volume 1, pages 321–328. doi: 10.1109/MDM.2014.46.

Xu, D.; Song, G.; Gao, P.; Cao, R.; Nie, X.; Xie, K. (2011). Transportation modes identification from mobile phone data using probabilistic models. In *International Conference on Advanced Data Mining and Applications*, pages 359–371. Springer.

Yang, Y.; Widhalm, P.; Athavale, S.; González, M. C. (2016). Mobility sequence extraction and labeling using sparse cell phone data. In *AAAI*, pages 4276–4277.

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

Zang, H.; Bolot, J. (2011). Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, pages 145–156. ACM.

Zannat, K. E.; Choudhury, C. F. (2019). Emerging big data sources for public transport planning: A systematic review on current state of art and future research directions. doi: 10.1007/s41745-019-00125-9.

Zhao, J.; Zhang, F.; Tu, L.; Xu, C.; Shen, D.; Tian, C.; Li, X.; Li, Z. (2017). Estimation of passenger route choice pattern using smart card data for complex metro systems. *IEEE Transactions on Intelligent Transportation Systems*, 18(4):790–801. doi: 10.1109/TITS.2016.2587864.

Glossary

ANPR Automatic number-plate recognition. 20

API Application Programming Interface. 41

billing data Data stored by the mobile operator for billing purposes, see also CDR. 15, 16

Bluetooth Standard for short-range wireless communication. 20

calibration Adjusting parameters of a model to ensure accurate results. 12, 26, 27, 33, 44, 59

CDR Call Detail Record. 16, 17, 29, 33, 46, 50, 54, 57

cellpath A list of cells which a user connected to during a trip. 29, 32, 33, 50, 57

cellular network data Records from the mobile network containing a timestamp and the cell which a user was connected to, see also CDR and xDR. 1–5, 15–18, 20, 23–26, 28–36, 39, 40, 42–52, 54, 55, 57–59

classification Mapping each observation to a label from a fixed set of categories; example: travel mode classification. 26, 27, 31, 32, 34, 37, 51, 52, 58

clustering Grouping observations into groups with similar observations (clusters). 27, 29, 34, 35, 42, 53, 55, 58

GIS Geographic Information System. 43, 47

GPS Global Positioning System. 11, 20, 36, 39, 49

GSM Global System for Mobile communications. 17

handover Switch between two cells during an ongoing phone call or data connection. 15, 16

HMM Hidden Markov Model. 27, 33, 35

lcoation update Records stored by a cellular network operator to support location management. 15, 16

link flow Number of traveller or vehicles on a given link in the transportation network. 9, 13, 28, 31, 32, 35, 37, 50

LTE Long-Term Evolution. 17

machine learning Algorithms learning autonomously from a set of training data. 26, 27, 31–35, 55, 57

modal split Share of trips made with each travel mode. 9, 31, 37, 38, 42, 51

OD-matrix Origin-destination travel demand matrix. 9, 30, 31, 35, 37

OD-pair Pair of two TAZs (origin and destination). 9, 11–13, 31, 32, 51, 52

PCA Principal component analysis. 27

PostGIS GIS extension for Postgres. 47

PostgreSQL Relational database management system. 47

precision Fraction of the expected observations among the detected observations. 37

Python Programming language. 47

QGIS GIS software. 48

R Programming language with focus on statistics. 48

recall Fraction of the detected observations among the expected observations. 37, 49

smart card data Data collected from a smart-card based Automatic Fare Collection (AFC) system used in public transit. 1–3, 5, 15, 18–20, 23, 25, 34, 35, 42–46, 48, 53, 55, 57, 58

supervised learning Algorithms learning autonomously from a set of training data with correct answer known. 27, 51, 58

TAZ Traffic Analysis Zone. 9, 30, 31, 35

test error Fraction of correctly classified observations. 37

travel behaviour Descriptions of the decision making processes how travellers choose to travel. 1, 8, 9, 12–14, 35, 42–45, 53, 55, 58, 59

UMTS Universal Mobile Telecommunications System. 17

unsupervised learning Algorithms learning autonomously from a set of training data without correct answer known. 27, 57

user equilibrium Route assignment in a transportation network such that no traveller can find a faster route. 13

Voronoi cell Polygon containing all points for which a given antenna is the closest; computed using Voronoi Tesselation (Aggarwal et al., 1989). 16, 50

xDR x-Detail Record. 16

Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-175347

FACULTY OF SCIENCE AND ENGINEERING

Linköping Studies in Science and Technology, Dissertation No. 2141, 2021
Department of Science and Technology

Linköping University
SE-581 83 Linköping, Sweden

www.liu.se