Analysis of Travel Patterns from Cellular Network Data

Nils Breyer
Analysis of Travel Patterns from Cellular Network Data

Nils Breyer
Analysis of Travel Patterns from Cellular Network Data

Nils Breyer

LIU-TEK-LIC 2019
ISBN 978-91-7685-055-8
ISSN 0280–7971

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

Printed by LiU Tryck, Linköping, Sweden 2019
Abstract

Traffic planners are facing a big challenge with an increasing demand for mobility and a need to drastically reduce the environmental impacts of the transportation system at the same time. The transportation system therefore needs to become more efficient, which requires a good understanding about the actual travel patterns. Data from travel surveys and traffic counts is expensive to collect and gives only limited insights on travel patterns. Cellular network data collected in the mobile operators infrastructure is a promising data source which can provide new ways of obtaining information relevant for traffic analysis. It can provide large-scale observations of travel patterns independent of the travel mode used and can be updated easier than other data sources. In order to use cellular network data for traffic analysis it needs to be filtered and processed in a way that preserves privacy of individuals and takes the low resolution of the data in space and time into account. The research of finding appropriate algorithms is ongoing and while substantial progress has been been achieved, there is a still a large potential for better algorithms and ways to evaluate them.

The aim of this thesis is to analyse the potential and limitations of using cellular network data for traffic analysis. In the three papers included in the thesis, contributions are made to the trip extraction, travel demand and route inference steps part of a data-driven traffic analysis processing chain. To analyse the performance of the proposed algorithms, a number of datasets from different cellular network operators are used. The results obtained using different algorithms are compared to each other as well as to other available data sources.

A main finding presented in this thesis is that large-scale cellular network data can be used in particular to infer travel demand. In a study of data for the municipality of Norrköping, the results from cellular network data resemble the travel demand model currently used by the municipality, while adding more details such as time profiles which are currently not available to traffic planners. However, it is found that all later traffic analysis results from cellular network data can differ to a large extend based on the choice of algorithm used for the first steps of data filtering and trip extraction. Particular difficulties occur with the detection of short trips (less than 2km) with a possible under-representation of these trips affecting the subsequent traffic analysis.
Acknowledgments

First and foremost, 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 long discussions and giving valuable feedback. I very much enjoy working together with you. This work would not have been possible without your commitment and efforts to get access to relevant data.

The research in this thesis has been to 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.

I want to thank my colleagues at the division of Communication and Transportation Systems (KTS). I really enjoy being part of a group this international, with different research fields and perspectives. Special thanks to Viveka Nilsson for all help with administration. In particular, I would like to thank the group of PhD students for great company, fun times and all mutual support. I gratefully acknowledge your appreciation of my home-baked cakes for PhD-fika. I also want to take the opportunity to thank my former colleagues and roommates at the University, Marcus Posada and Anna Vasilevskaya for the amazing time together.

Last but not least, I would like to thank all my friends, my sister Annalena and my dear parents Marita and Gerd-Herbert for their unconditional support.

Norrköping, May 2019

Nils Breyer
# Contents

Abstract iii

Acknowledgments v

1 Introduction 1  
1.1 Motivation 1  
1.2 Aim and scope 2  
1.3 Outline 3

2 Traffic analysis using cellular network data 5  
2.1 Traffic analysis 5  
2.2 Data-driven traffic analysis 7  
2.3 Cellular network data 8  
2.4 Previous research 10  
2.5 Research gaps 13

3 Method and contributions 15  
3.1 Method 15  
3.2 Contributions 17  
3.3 Summary of the included papers 17

4 Conclusions and Future research 23

Bibliography 25

Abbreviations 31

Paper I 33

Paper II 41

Paper III 77
Chapter 1

Introduction

The demand for mobility keeps increasing in many countries and cities, while at the same time environmental impacts caused by the traffic system need to be reduced drastically in order to meet global climate goals. Traffic planners are hence facing a huge challenge to make the traffic system more efficient. To make well-informed decisions when developing the traffic system, a detailed understanding of human mobility is needed. Comprehensive information on travel patterns and the actual travel demand is today difficult to obtain. Cellular network data collected using the mobile operators existing infrastructure is a promising new data source which can help to make more informed decisions in traffic planning and traffic management. The goal of this thesis is to contribute to the understanding of what is needed to make cellular network data usable in a traffic analysis context and to show the potential as well as limitations of this data source.

1.1 Motivation

Today, common ways to gather information on travel patterns are travel surveys and traffic counts. Both are expensive to conduct and give only limited insights on mobility patterns. Travel surveys are also facing challenges with under-reporting and decreasing response rates (Prelipcean et al., 2015; Schulz et al., 2016). Traffic models are used to estimate travel demand, mode choice and route choice based on a number of behavioural assumptions. While these models can
provide a reasonable estimate, they only offer a limited level of detail and as the assumptions do not always hold in reality, there is a high level of uncertainty in the resulting estimates.

Cellular network data is already collected by mobile network operators today for different purposes. For each user the antenna which the user is connected to together with a timestamp is saved on certain events, for example when the user makes a call. As a large-scale data source, cellular network data could provide a more comprehensive understanding of travel patterns. Given the ubiquity of mobile devices, trips with all travel modes can be observed, while traffic counts only cover one travel mode, such as the number of vehicles using a specific road or passenger counts in public transport. The data is also easy to update and could potentially even be processed in real-time. This makes it not only usable for long-term traffic planning, but also short-term traffic management applications.

To make cellular network data usable for traffic analysis, the raw data needs to be processed and aggregated. Challenges when processing the data occur as cellular network data has often a low resolution in space and time. Unlike other data source, the resolution also varies in different situations and the data contains noise which is not related to trips made between places. Besides the need for new methods to process the data to make it usable for traffic analysis, there is also a need to find ways to evaluate these methods and the results they produce. This is necessary both to compare different methods as well as to better understand the quality of results obtained from cellular network data compared to other data sources.

1.2 Aim and scope

The aim of this thesis is to improve the understanding of the potential and limitations of using cellular network data for traffic analysis. This includes to propose different data-driven methods to understand travel patterns from cellular network data and to investigate differences between the results they produce. To provide realistic results, the methods are applied to data obtained from existing cellular networks. By analyzing the results for different datasets, comparing different methods against each other and comparing results with other existing data sources, the problem of using cellular network data to understand travel patterns is analysed from different perspectives.
The following main research questions have been guidance for the work within the thesis:

- How can cellular network data be processed algorithmically in order to estimate quantities relevant for traffic analysis? What is the impact of using different algorithms on the results?
- How can the quality of the results obtained from cellular network data be evaluated? How can the results produced be compared with other data sources?
- What potential and obstacles of using cellular network data for traffic analysis can be identified from applying the algorithms to dataset collected from existing cellular networks?

The focus of the thesis is the data processing from cellular network data in a traffic analysis context and the evaluation of the results. The focus is not the data collection and cellular network signalling aspects. The term cellular network data here refers only to updates that occur in the network of the mobile operator containing a timestamp and the antenna which a user was connected to. Other types of data which occur in cellular networks, such as measurement reports are not used. While the thesis outlines a process for a fully data-driven traffic analysis pipeline, some parts of the process are left for further research or are only partly covered. This includes for example the distinction of travel patterns by travel mode. The methods presented in this thesis are fully data-driven and based on cellular network data. The data fusion with other sources and a potential integration of the data into traffic models are not covered in this thesis.

1.3 Outline

The remainder of the thesis is organized as follows. Chapter 2 introduces the background as well as the field of using cellular network data for traffic analysis including previous research and the main research gaps. Chapter 3 introduces the research method and describes the contributions made in the included research papers as well as how the papers are interlinked. The main conclusions and topics left for future research are discussed in Chapter 4. The three papers which are part of this thesis are included in the appendix.
Chapter 2

Traffic analysis using cellular network data

To give an overview of the needs and current practices used by traffic planners to understand travel patterns, this chapter will provide a summary of the models and data sources currently in use as well as emerging new data sources. Being the data source used throughout this thesis, cellular network data is introduced in detail. The research which has been done using cellular network data for traffic analysis is summarized and the remaining research gaps in the field are discussed.

2.1 Traffic analysis

The term traffic analysis here refers to the part of traffic engineering which deals with describing and understanding historic, current and future traffic patterns based on models and data. Traffic analysis can be done both on a microscopic level to understand the behaviour of individual vehicles or humans in a small part of the transportation network or on a macroscopic level for a region or city, where the traffic is described in terms of total flows rather than individual vehicles (Treiber and Kesting, 2013). Analysing traffic patterns is elementary to be able to make improvements to the traffic system. The understanding of when and where travel demand occurs, which travel modes and routes travellers choose is elementary for both long-term traffic planning and short-term traffic management.
A common way to estimate aggregated travel patterns is to use census data and a description of the transportation infrastructure as an input to a traffic model, which then estimates travel patterns based on a number of behavioral assumptions. The standard procedure is to follow the so-called four-step model, which is being described for example in McNally (2007). The first step of the four-step model, as shown in Figure 2.1, is the trip generation, which estimates the total travel demand in each Traffic Analysis Zone (TAZ), typically using demographic data, landuse or statistics on the number of workplaces and facilities in each area. In the second step, the trip distribution, an OD-matrix is calculated such that the total demand calculated previously is distributed among the available OD-pairs using their travel times and under the assumption that people prefer to visit facilities which are faster to reach over those taking longer time to travel to. The demand in each OD-pair is then distributed among the available travel modes and finally, in step 4, among different routes available for the travel mode. This procedure allows to estimate the flows in the transportation network as, for example, the number of travellers on a given road stretch or the number of passengers on a public transit line. As the number of travellers on a link affects the travel times, which in turn are input to the trip distribution step, a feedback loop is often used from the fourth to the second step. The steps are then repeated until convergence.

The four-step model describes the general and widely accepted procedure used in traffic modelling. The specific models used for each step can vary in practice. A common model used for the trip distribution step is the Gravity model (Erlander and Stewart, 1990) which distributes the travel demand under the assumption that the number of trips in an OD-pair decays with increasing travel time (or generalized cost). Travel mode choice is often modelled using a Logit model (Wen and Koppelman, 2001) and in the network assignment step a procedure to find a user equilibrium state, that is a state were all used routes in an OD-pair share the same travel time, can be used.
2.2. Data-driven traffic analysis

for vehicular traffic (Patriksson, 2015). A common assumption made in these models is that most people want to minimize their travel time and cost. Further, it is assumed that travellers have perfect knowledge on the actual travel time and cost. While the assumptions in the models might not hold perfectly in reality, they allow to model and analyse traffic flows from aggregated demographic data and a description of the transportation network. For the road network that would be the links and their travel time functions. Travel surveys and traffic counts provide observations which can be used in order to calibrate parameters and validate.

Observations from travel surveys and traffic counts are expensive to obtain and therefore limited in their availability. Comprehensive travel surveys are usually only carried out infrequently in intervals of several years and suffer from decreasing response rates (Schulz et al., 2016). Outdated, non-representative and incomplete data makes the output of traffic models as used in the four-step model unreliable. Traffic counts require expensive equipment (or labour) and can therefore only be collected at a few selected places in the traffic network. The need for the input data limits the use of traffic models to regions where this data is available. In developing countries for example the lack of adequate data can make it impossible to use traffic models to analyse traffic. A challenge is also to update traffic models to the changing behaviour of travelers, for example due to the rise of new mobility services.

2.2 Data-driven traffic analysis

With new technology available and mobile devices being widespread, new opportunities open to gain insights on human mobility using new data sources. While traffic analysis traditionally models travel patterns based on statistics and assumptions, more data-driven approaches are on the rise which make use of large-scale observations of trips.

Besides data from mobile phone operators, which will be covered in the remainder of this thesis, there are a number of other emerging data sources which can be used for traffic analysis. One of these are Global Positioning System (GPS) tracks recorded on mobile devices using the global satellite based positioning system. GPS allows to record tracks of high spatial accuracy and temporal resolution (Bar-
Large-scale GPS datasets are collected and owned by companies providing navigation apps. Similarly, also a fleet of probe vehicles can be used to collect GPS tracks (Hofleitner et al., 2012). While these datasets can be used to estimate travel times and detect traffic jams in real-time, they are less useful to estimate travel demand due to possible bias in the observations. Navigation apps for example are mostly used by car drivers for longer, unknown routes rather then the everyday commuting trip with public transport. Recently, efforts are made as well to replace travel surveys on paper with GPS supported travel diaries (Prelipcean et al., 2015) which could increase data quality and lower costs for carrying out travel surveys.

Other possible sources are to use Bluetooth or WiFi access points to count devices passing by (Barceló Bugeda et al., 2012). For public transport, smartcards which are used to buy tickets can be used to count travellers when entering the public transport system (Anda et al., 2017). Some social networks also allow geo-tagging of public user posts. Even though this data is rarely representative for a whole population, it can contribute to the understanding of human mobility (Cho et al., 2011).

New large-scale data sources open up possibilities to understand mobility patterns from actual observations. On the other hand, there are a number of things which need to be considered such as varying data quality, problems with the representativeness and privacy considerations. Despite their large-scale nature, even data-driven approaches to traffic analysis need to make assumptions in order to describe traffic patterns for a whole population, as the data is never complete, as well as to be able to predict future traffic.

### 2.3 Cellular network data

The data source which is used in this thesis is cellular network data. This term refers to data which occurs in cellular networks and is collected by network operators for different reasons. Following the classification of different types of cellular network data in Gundlegård (2018), this can include billing data, location updates, handovers, measurement reports and dedicated location data.

Billing data is collected for example when a phone is used to make a phone call. Datasets where each event corresponds to a phone call or SMS are also referred to as *Call Detail Records* (CDR). If addi-
2.3. Cellular network data

Table 2.1: An example of an artificial cellular network dataset.

<table>
<thead>
<tr>
<th>User ID</th>
<th>Timestamp</th>
<th>Antenna ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2018-10-01 06:50:00</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2018-10-01 08:10:00</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2018-10-01 08:20:00</td>
<td>2</td>
</tr>
</tbody>
</table>

tional records are saved on other events as ongoing data connections, the term \textit{x-Detail Records} (xDR) is used. Location updates and handovers are related to a mobile device’s change of location area respectively antenna which the device is connected to (Saifullah et al., 2012). The difference is that location areas consist of multiple antennas and therefore location updates are triggered less often than handovers. Billing data, location updates and handovers can be represented by a user id, timestamp and the antenna ID of the antenna which the user is connected to, as in the example in Table 2.1. Together with the known position of the antennas an approximate area describing the possible position of the user can be estimated. The estimated position can have an uncertainty of up to several kilometers depending on the density of antennas in the area.

While the above datatypes give an approximate position based on the coverage of the antenna, measurement reports and dedicated location data can give more accurate position data using signal strength and round-trip-time measurements, which can be comparable to GPS data. Data collection efforts and privacy implications are therefore much higher for these types of data. In this thesis no measurement reports or dedicated location data have been used. The term cellular network data in the remainder of the thesis refers to data of the form shown in Table 2.1 together with the location of the the antennas and potentially other metadata about the antennas which can be used to estimate their coverage area.

Cellular network data can provide large-scale observations of travel patterns for the subscribers of an operator, which often is a significant share of the population. As mobile devices are ubiquitous today, movements with all travel modes can be observed. As it makes use of the existing infrastructure, there is no need to install additional equipment. This also makes it possible to collect updated data regularly and possibly even in real-time.
Despite these advantages, there are a number of challenges which occur when using cellular network data for traffic analysis. As the cellular network is not designed as a positioning system, the connection to a specific antenna gives only a proxy of the position. Depending on many factors and the density of antennas around, the coverage range of an antenna can be several kilometers (Bhaskaran et al., 2003). This implies also that the accuracy of the estimated position varies between different areas. Not only the spatial resolution varies, also the resolution in time can be different depending on how and which the data is collected. In case of CDR data, for example, the time resolution directly depends on the user’s phone call frequency. Other datasets might contain periodic updates which occur with a fixed time interval. A switch of antennas can be caused by many reasons, which not necessarily must be a movement of the user between places. A switch can be caused due to the network balancing load between different antennas or a change in weather conditions. Additionally, modern phones use different networks such as the Global System for Mobile communications (GSM), Universal Mobile Telecommunications System (UMTS) and Long-Term Evolution (LTE). Antennas of different networks can overlap in coverage and devices can switch between networks at any time.

Information on when and which places individuals visit is sensitive data and therefore needs to be protected. De Montjoye et al. (2018) have categorized the different approaches which are used to protect privacy, while still being able to use the data to extract anonymized and aggregated data, such as travel patterns. Two of these concepts are used in the work within this thesis. The first is “Limited release”, where the raw data is manipulated using re-hashing of user-ids or obfuscated in other ways. The second is “Remote access”, where the raw data stays in the premises of the operator at all times and only high-level aggregated data is being exported for later analysis.

Previous research

Research with the aim to make cellular network data usable to understand human mobility has been ongoing for a decade, with major contributions by Caceres et al. (2007); Gonzalez et al. (2008); Calabrese et al. (2011); Alexander et al. (2015) among many others. There is a wide range of use cases of the data in the area of traffic analysis.
2.4. Previous research

One of these is to better understand human mobility in general using cellular network data as a source of large-scale observations. Gonzalez et al. (2008) have demonstrated the possibilities in this area as early as 2008. More recently, ways to understand human mobility behaviour in terms of general statistics and predictability have been presented by Cuttone et al. (2018); Barbosa et al. (2018). While a deeper understanding of human behaviour is interesting itself, it is also gaining more practical importance. With the appearance of new mobility services and types of vehicles, a deeper understanding of mobility behaviour could allow to predict consequences of new forms of mobility better.

Caceres et al. (2007) were among the first to show the possibility to use cellular network data as a source of observations that can be used to describe the actual travel demand. As data on travel demand is otherwise difficult to obtain, it could be used to make models and forecasts more reliable. This would allow to make more informed decisions on how the traffic infrastructure should be developed (Becker et al., 2011). An example is the planning of changes to the road and railway infrastructure, but also the public transport network and timetables. Collecting data before and after changes to the infrastructure allows to better analyse the effects which a specific change had. In regions where the necessary data for traditional traffic models, like travel surveys and traffic counts, is not available, cellular network data could be the very first way to get an overview of the travel patterns.

With the possibility to process cellular network data in real-time, it is not only useful for long-term traffic planning, but could also be used for short-term traffic management as demonstrated by Janecek et al. (2015). Using the cellular network as a real-time sensor could allow to detect unusual high travel demand faster and taking appropriate countermeasures. Large-scale observations could also be used to train and improve traffic predictions. Cellular network data could also be used to understand movements in case of big events and special situations like catastrophes (Marques-Neto et al., 2018; Calabrese et al., 2010). Data covering countries or even larger areas are of use to better understand epidemics, migration and mobility connected to tourism (Barbosa et al., 2018).

To make cellular network data usable for traffic analysis, a number of processing steps are necessary. As the raw data contains events which not necessarily are connected to movements, most studies im-
implement some kind of trip detection as one of the first steps. The goal is to identify the sections in the data which correspond to trips where a user moved between two locations. The main challenge is to distinguish movements of the user from antenna switches made for other reasons. A common approach is to identify standstill sections in the data according to some thresholds, as for example done by Calabrese et al. (2010). Alexander et al. (2015) have used recurring places like home and work to improve trip detection. Another approach to detect stops and trips using clustering has been presented by Widalham et al. (2015).

Using cellular network data to infer an OD-matrix describing the travel demand has been and is a field of active research. Early contributions have been made by Caceres et al. (2007) and Calabrese et al. (2011) among others. Recently, efforts are made to split the travel demand estimated by characteristics as the travel mode (Bachir et al., 2019; Huang et al., 2019) and trip purpose (Toole et al., 2015). Work has also been done finding methods to recover route choices from cellular network data, even if this field has gotten less attention then travel demand estimation. Contributions in the area have been made by Gundlegård and Karlsson (2009); Song et al. (2017).

As the methods to process the data itself differ, also the methods to compare and validate the results are varying among different studies. Only few studies compare individual trips extracted as done by Fillekes (2014). In most studies, results are compared on an aggregated level for a large population. This is commonly done either by comparing general statistics as for example the number of trips made or the trip length distribution (Liu et al., 2014) or by analysing the correlation to other available data (Batran et al., 2018).

It is apparent that traditional traffic analysis methods need to change in order to take full advantage of the potential of new large-scale data sources like cellular network data. While many studies in the research field use a largely data-driven approach (Anda et al., 2017), this might not be the best way in practice for many applications. The central open question is how cellular network data is best used for different traffic analysis applications and how it is best combined with traffic models. Some studies, as for example Huang et al. (2018), propose to use it to improve the calibration of existing models and others suggest same processing of the data which then can be used as input to traffic models (Tolouei et al., 2017). Wismans et al. (2018) presented a method where the output of a traffic model
is ex-post combined and augmented using cellular network data. Connected to this question is which role traditional data sources as travel surveys and traffic counts will play in future traffic analysis methods. Calabrese et al. (2013); Von Mörner (2017) argue that these sources will be needed even if large-scale cellular network data is available, however the need for these expensive data sources could possibly be reduced.

2.5 Research gaps

Despite the research which already has been done regarding the use of cellular network data for traffic analysis, many open questions remain. The possible applications are broad and many different methods might be used to process the data. Major research gaps in the field continue to exist within at least four different areas: Data collection, Data processing algorithms, Evaluation of results and the Integration with models. This thesis aims to makes contributions to these gaps with a special focus on data processing algorithms and the evaluation of results.

Regarding the data collection of cellular network data, the understanding of how the characteristics of different cellular network data datasets affect traffic analysis results is incomplete. Different datasets can vary in their spatial and temporal resolution and it has not been quantified how different characteristics could affect the results based on the data. The reliability of the data collection of cellular network data is often unknown and could be analysed more systematically. Many studies are using the Voronoi cells as a simple estimation of the coverage area. The effect of using more accurate representations of the coverage areas has not been fully investigated.

The data processing algorithms are central to transform cellular network data into information usable for traffic analysis. A large number of methods have been tested, each of them having certain limitations. Finding the most appropriate method for a given dataset and application however remains a challenge. More systematic comparisons of different methods are necessary to understand their performance. On the way to practical use for traffic analysis, a number of problems are not fully solved. Particularly challenging aspects, as the classification of travel modes or the scaling of travel demand to the whole population have not yet gotten the necessary attention. Many
studies focus on few processing steps and types of results, while for practical use several might need to be combined. The effects of combining different processing steps and methods needs to be investigated more. Also, processing data in real-time and for larger areas poses new constraints on computation time and might require dedicated methods. Finally, the different concepts to preserve privacy while processing the data described by de Montjoye et al. (2018) need to be developed further.

Due to the lack of a ground truth for most results, the evaluation method for the results obtained from cellular network data is not obvious. Validation and comparison can be done on different levels, using different comparison datasets and using different metrics measuring similarity. Finding appropriate evaluation methods and metrics to understand the quality of the results is important to compare different methods. Quantifiable indicators of the quality of the results are also necessary to show the usability of the data for practical traffic planning applications and the limitations of cellular network data. While aggregated statistics can be compared easily, it is difficult to measure for example the similarity of travel patterns represented in an OD-matrix taking into account the spatial structure.

Data-driven methods using cellular network data can potentially provide a much better understanding of the actual travel patterns. However, they will never provide observations of all movements and therefore assumptions are still necessary to project the observations to the whole population. Traffic models could be used for that purpose. A major research gap exists, regarding how data from cellular network data can be combined with traffic models as well as other possible data sources in the best possible way. Very little research has been done in this area and there is a chance that using the strengths of different data sources and models could allow for more reliable traffic analysis methods.
Chapter 3
Method and contributions

This chapter summarizes the work which has been done within this thesis. First, the general research approach is described. The interconnections between the included papers are illustrated and related to the research approach. The contributions of the thesis to the existing research gaps are summarized. Finally, a summary of the three included papers is given.

3.1 Method

The work conducted within this thesis can be related to the general data-driven traffic estimation pipeline shown in Figure 3.1. The first version of this processing pipeline has been presented in Gundlegård

![Figure 3.1: Processing pipeline for data-driven traffic analysis form cellular network data](image)
Chapter 3. Method and contributions

et al. (2016), with contributions by the author of this thesis. The pipeline is inspired by the traditional four-step model (see Chapter 2.1), but instead of using traffic models it is purely data-driven and uses cellular network data as input. The papers included in this thesis make contributions to different parts of the pipeline (see Chapter 3.3).

The first step of the pipeline is the trip extraction step, which has the goal to distinguish time periods of movements and standstill in the raw cellular network data. Only the movements, that is trips with a given start and end time and a start and end position defined by given antennas, are of interest in the context of traffic analysis. Compared to the traditional four-step model, this step combines both the trip generation and trip distribution steps. While trips extracted from cellular network data have similarities with trips reported by participants in a travel survey, there are a number of differences due to the different way of data collection (Tolouei et al., 2017). Besides the much larger sample size in cellular network datasets, more metadata about trips made, such as the used travel mode, can be collected in travel surveys by directly posing questions, while from cellular network only the time and locations can be estimated directly.

The second step of the processing pipeline is the mode classification with the goal to identify which travel modes have been used for the trips detected. This step is not covered in this thesis and identified as a matter of further research. The trips detected from cellular network data are then used to estimate travel demand aggregated and scaled to the whole population. The route estimation also uses the trips as input and tries to estimate possible routes used in each origin-destination pair and their popularity. The final flow distribution step is comparable to the network assignment step in the four-step model. It assigns the travel demand to the transportation network using the routes estimated, which yields flow for each link in the transportation network.

The algorithms proposed in the included papers have been implemented and tested using different datasets. As cellular network data is a large-scale data source, one of the challenges is to develop and implement algorithms such that they are computationally efficient. To achieve that, a combination of code in Python and a Postgres database with the PostGIS extension providing a Geographic Information System (GIS) is used. The database allows to store and query data fast for processing and analysis. All processing algorithms have been parallelized. A significant part of the work has been to analyse...
3.2 Contributions

The papers included in this thesis make contributions to different parts of the processing pipeline (see Figure 3.1) and proposes algorithms to process the data and methods to evaluate the results. The main contributions of the thesis are:

- three algorithms to extract trips from cellular network data distinguishing movements and standstill using different approaches (Paper I, Paper II),
- an analysis of the effects of the type of cellular network data on trip extraction (Paper I),
- the evaluation of the quality of trips extracted by comparing trip-by-trip (Paper I, Paper II),
- a method to infer travel demand using previously extracted trips (Paper II),
- an analysis of the effect of the trip extraction method on travel demand estimation (Paper II),
- algorithms to estimate routes on the road network from cell-paths (Paper III),
- an analysis of the impact of the route estimation algorithms on the network assignment (Paper III).

3.3 Summary of the included papers

The included papers and their contributions can be placed along the processing pipeline (see Figure 3.1). As the processing steps build up on each other, the included papers have been numbered in order to
follow the order in the processing pipeline, rather than their date of publication. Paper I focuses on the trip extraction as the first step of the process. Paper II focuses on the travel demand estimation step, while also making further contributions to the trip extraction step. Paper III makes contributions to the route estimation and flow distribution steps of the process. In the following, the three included papers are summarized.

**Paper I: Trip extraction for traffic analysis using cellular network data**

In Paper I, two trip extraction algorithms are presented: The first algorithm is using important places (Home/Work) and the second is based on movement indicators (Efficiency). Their performance is compared using two different datasets. A small dataset collected in Sweden is used to compare the trips extracted from cellular network data with trips collected using Google location history, which uses GPS data, from the same mobiles. By applying the same methods to a large-scale dataset for Senegal it is shown how the choice of the trip extraction method impacts the detected trips, which is important to understand when using the extracted trips for further traffic analysis applications.

Using the two datasets it is illustrated that the choice of the trip extraction method tends to be more important when the sampling rate of the raw data is low. Despite the much lower location sampling rate in the cellular network data, the results are promising and show that most trips found from GPS can also be detected from cellular network data. For example, for the Efficiency algorithm, a recall of 67% is achieved.

Paper I is co-authored with David Gundlegård, Clas Rydergren and Johan Bäckman. The author of the thesis has contributed as main author and major work with the development of algorithms and analysis of the results.

Paper I has been presented at


Parts of Paper I have been presented at
3.3. Summary of the included papers

- the Swedish transportation research conference 2017, Stockholm, Sweden and

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

This paper shifts focus towards the travel demand estimation step of the processing pipeline. A data-driven travel demand estimation method is presented which consists of a trip extraction step building upon the work presented in Paper I, extended by a travel demand estimation step. In a comparative analysis the potential and limitations of the method are investigated. To find out which types of trips can be extracted from cellular network data, the same small scale cellular network dataset as in Paper I collected from 20 mobile phones together with GPS tracks collected on the same device is used. A large-scale dataset of cellular network data from a Swedish operator for the city of Norrköping is then used 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 transit 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 transit, 74% for car but only 53% for bicycle and walking. The correlation to the urban travel demand model used by the municipality 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. Using two different trip extraction methods, systematic differences in the resulting travel demand can be observed, which highlights that the choice of the trip extraction method is crucial for the travel demand estimation.

Paper II is co-authored with David Gundlegård and Clas Rydergren. The author of the thesis has contributed as main author and major work with the development of algorithms and analysis of the results.
Paper II is a working paper.

Parts of Paper II have been presented at


The content of Paper II has been accepted as poster at

- Netmob 2019, Oxford, UK.

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

This paper makes contributions to the route estimation and flow distribution steps of the processing pipeline. Different algorithms to estimate a route on the road network for a given cellpath are presented: Direct routing (shortest path between origin and destination), Waypoint Routing (route has to go through all Voronoi cells in the cellpath), Magnetic Routing (route follows the cellpath using modified link costs) and Magnetic Waypoint Routing (combines the previous two algorithms). The route estimation methods are compared regarding the characteristics of the individual routes. To investigate the effect of the different route estimation methods when used in a network loading for a larger city, a large-scale CDR dataset for Dakar, Senegal is used. The results show that the choice of the route estimation method has a significant impact on the resulting link flows.

Paper III is co-authored with David Gundlegård and Clas Rydergren. The author of the thesis has contributed as main author and major work with the development of algorithms and analysis of the results.

Paper III has been published in


Paper III has been presented at

- MobileTartu 2016, Tartu, Estonia.
3.3. Summary of the included papers

Parts of Paper III have been presented at

- *Netmob 2017*, Milan, Italy and
- the *Swedish transportation research conference 2016*, Lund, Sweden.
Chapter 4
Conclusions and Future research

Cellular network data is an emerging data source which could allow traffic planners to make more well-informed decisions when developing the transportation network. It is a large-scale data source which makes it possible to understand travel patterns based on many observation providing greater detail than with travel surveys. Further, it is relatively easy to obtain updated data regularly, possibly even in real-time. In contrast to traffic counts it allows to observe mobility patterns with all travel modes.

In order to make cellular network data usable for traffic analysis, extensive data processing is necessary. This thesis makes contributions to several parts of a possible data-driven process for traffic analysis from cellular network data. An elementary step necessary to use the data in a traffic analysis context is to identify movements from the raw data in order to extract trips made between different places. A number of different algorithms to extract trips from cellular network data are proposed in this thesis. Further, a method to infer travel demand using previously extracted trips is described. Finally, algorithms to estimate routes from cellular network data are proposed. Besides the algorithms itself, contributions are made by proposing adequate validation and comparison methods in order to evaluate the performance of the algorithms.

All of the proposed algorithms have been applied to cellular network data collected from existing networks in order to understand
the potential as well as practical limitations of the proposed methods and the available data. The results show that due to the low resolution and high level of noise in cellular network data, the filtering and processing methods have a significant impact on the results. The method used for the first processing step, the trip extraction, for example, impacts all later results significantly and needs to be chosen carefully. The trip-by-trip validation of trips extracted from cellular network data presented in this thesis showed that, in particular, short walking and cycling trips are difficult to capture and therefore likely to be under-represented.

While the resolution of the data set limits for the results that can be obtained and the privacy of individuals needs to be preserved, the research presented in this thesis shows that cellular network data has great potential to provide a more detailed overview of aggregated travel patterns independent of the travel mode. It allows to get insights for each individual origin-destination pair, including the time profile of the travel demand for a specific pair of zones. Travel surveys usually contain too few observations to obtain this level of detail.

There is still a large potential for further research regarding the use of cellular network data for traffic analysis. Possibilities of breaking down travel patterns by different modes, activities or groups of travellers have not fully been investigated and more work is required to ensure that the results are representative for the whole population. A challenge with ensuring the quality of the results is the absence of large-scale ground truth data. To compare data obtained from cellular network data to existing data sources, appropriate comparison metrics are necessary. While some ways of comparison are presented in this thesis, there is still a need for additional comparison methods and better metrics.

Further research is necessary to investigate the possibilities to use cellular network data in real-time and for traffic predictions, which could open for uses within traffic management. This thesis focuses on data-driven methods relying on cellular network data as the main data source. The possibilities to combine cellular network data with other data sources as for example sensors, ticket data, census data, GPS tracks or travel surveys are not yet fully explored. Besides the use of other data sources, also the integration of travel patterns obtained from cellular network data with traffic models is a promising approach.


Hofleitner, A.; Herring, R.; Abbeel, P.; Bayen, A. (2012). Learning the dynamics of arterial traffic from probe data using a dynamic


Abbreviations

CDR  Call Detail Record. 8, 10, 20

cellpath A list of cells (antennas) which a user connected to during a trip. 17, 20

cellular network data Records from the mobile network containing a timestamp and the antenna which a user was connected to, see also xDR. iii, 1–3, 5, 8–14, 16–19, 23, 24

four-step model Traffic model consisting of trip generation, trip distribution, mode choice and route choice steps (McNally, 2007). 6, 7, 16

GIS  Geographic Information System. 16

Google location history Service to track movements using a mobile phone’s sensors including GPS. 18

GPS Global Positioning System. 7–9, 18, 19

Gravity model Model to estimate travel demand for each OD-pair given an attraction of each zone (Evans, 1973). 6

GSM Global System for Mobile communications. 10

Logit model Model to estimate choice probabilities given the utility of different alternatives Wen and Koppelman (2001). 6

LTE Long-Term Evolution. 10

OD-matrix Origin-destination travel demand matrix. 6, 12, 14

OD-pair Pair of two TAZs (origin and destination). 6
Abbreviations

**PostGIS** GIS extension for Postgres. 16

**Postgres** Relational database management system. 16

**Python** Programming language. 16

**QGIS** GIS software. 17

**R** Programming language with focus on statistics. 17

**recall** Fraction of the detected trips among the actual trips. 18, 19

**TAZ** Traffic Analysis Zone. 6

**UMTS** Universal Mobile Telecommunications System. 10

**user equilibrium** Route assignment in a transportation network such that no traveller can find a faster route. 6

**Voronoi cell** Polygon containing all points for which a given antenna is the closest; computed using Voronoi Tesselation (Aggarwal et al., 1989). 13, 20

**xDR** x-Detail Record. 9
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-157139