PhD-FSTC-2018-14
The Faculty of Sciences, Technology and Communication

DISSERTATION

Defence held on 07/02/2018 in Luxembourg

to obtain the degree of

DOCTEUR DE L'UNIVERSITÉ DU LUXEMBOURG

EN INFORMATIQUE

by

Thierry DERRMANN
Born on 24 June 1987 in Dudelange (Luxembourg)

MOBILE NETWORK DATA ANALYTICS FOR INTELLIGENT TRANSPORTATION SYSTEMS

Dissertation defence committee

Prof. Dr. Thomas Engel, dissertation supervisor
Professor, University of Luxembourg

Dr. Raphaël Frank
Research Scientist, University of Luxembourg

A-Prof. Dr. Francesco Viti, Chairman
Associate professor, University of Luxembourg

Dr. Marco Fiore
Researcher, National Research Council of Italy

Prof. Dr. Falko Dressler, Vice Chairman
Professor, University of Paderborn

May 6, 2018

Acknowledgements

This work was conducted at the SECAN-Lab of the University of Luxembourg's Interdisciplinary Centre for Security Reliability and Trust (SnT). The project was funded by the Luxembourg Fonds National de la Recherche (FNR).

First and foremost, I would like to thank Prof. Dr. Thomas Engel for letting me conduct my research at SECAN-Lab, allowing me to freely explore, identify and investigate my topics of interest, while always being supportive.

I would also like to thank Dr. Raphaël Frank for his continuous advice, motivation, and the much-needed criticism of my work, but also for showing me around campus at UCLA.

I also want to thank Prof. Dr. Falko Dressler for taking the time to supervise my progress and for the excellent work done by him and his team at the University of Paderborn, especially Dr. Christoph Sommer and Florian Hagenauer, co-creators of Veins and VeinsLTE. Without their works, this dissertation would not have been possible.

Further, I want to thank Prof. Dr. Francesco Viti, who supported me in my venture into the field of transportation, giving me invaluable pointers to the right models and literature, especially in traffic flow theory, and opportunities for joint publications at transportation conferences. I also want to thank his MobiLab team members Guido Cantelmo and Dr. Marco Rinaldi for their help with the transportation-related topics in this thesis.

Finally, I would like to thank all my colleagues at SECAN-Lab, MobiLab and at SnT who made work and conferences so much more enjoyable. Most of all, I want to thank my family and friends, who have always been supportive of my efforts and understanding when things were not going as planned.

Abstract

In this dissertation, we explore how the interplay between transportation and mobile networks manifests itself in mobile network billing and signaling data, and we show how to use this data to estimate different transportation supply and demand models.

To perform the necessary simulation studies for this dissertation, we present a simulation scenario of Luxembourg, which allows the simulation of vehicular Long-Term Evolution (LTE) connectivity with realistic mobility.

We first focus on modeling travel time from Cell Dwell Time (CDT), and show – on a synthetic data set– that we can achieve a prediction Mean Absolute Percentage Error (MAPE) below 12%. We also encounter proportionality between the square of the mean CDT and the number of handovers in the system, which we confirmed in the aforementioned simulation scenario. This motivated our later studies of traffic state models generated from mobile network data.

We also consider mobile network data for supporting synthetic population generation and demand estimation. In a study on Call Detail Records (CDR) data from Senegal, we estimate CDT distributions to allow generating the duration of user activities, and validate them at a large scale against a data set from China. In a different study, we show how mobile network signaling data can be used for initializing the seed Origin-Destination (O-D) matrix in demand estimation schemes, and show that it increases the rate of convergence.

Finally, we address the traffic state estimation problem, by showing how handovers can be used as a proxy metric for flows in the underlying urban road network. Using a traffic flow theory model, we show that clusters of mobile network cells behave characteristically, and with this model we reach a MAPE of 11.1% with respect to floating-car data as ground truth. The presented model can be used in regions without traffic counting infrastructure, or complement existing traffic state estimation systems.

Contents

Acronyms

AP Access Point.

ATIS Advanced Traveller Information System.

BC Bluetooth Classic.

BLE Bluetooth Low Energy.

C-ITS Cooperative Intelligent Transportation Systems.

CDR Call Detail Records.

CDT Cell Dwell Time.

CGF Charging Gateway Function.

DARC Data Radio Channel.

DSRC Dedicated Short Range Communications.

FCD Floating-Car Data.

FD Fundamental Diagram of Traffic Flow.

GGSN Gateway GPRS Support Node.

GMSC Gateway Mobile Switching Center.

GPS Global Positioning System.

GSM Global System for Mobile Communications.

GTFS General Transit Feed Specification.

GTFS-RT General Transit Feed Specification: Real-Time.

ICT Information and Communication Technology.

IMEI International Mobile Equipment Identifier.

IMSI International Mobile Subscriber Identifier.

ITS Intelligent Transportation System.

LTE Long-Term Evolution.

LWR Lighthill-Whitham-Richards model.

MAC Medium Access Control.

MAPE Mean Absolute Percentage Error.

MFD Macroscopic Fundamental Diagram.

MIMO Multiple-Input and Multiple-Output.

MME Mobility Management Entity.

MNO Mobile Network Operator.

MSC Mobile Switching Center.

NFD Network Fundamental Diagram.

O-D Origin-Destination.

PDN-GW Packet Data Network Gateway.

RAN Radio Access Network.

RNC Radio Network Controller.

RSSI Received Signal Strength Indicator.

RSU Road-Side Unit.

S-GW Serving Gateway.

SGSN Serving GPRS Support Node.

SNR Signal-to-Noise Ratio.

SSID Service Set Identifier.

TDM Transportation Demand Management.

TETRA Terrestrial Trunked Radio.

TIS Traffic Information System.

UMTS Universal Mobile Telecommunications System.

V2I Vehicle-to-Infrastructure Communication.

V2V Vehicle-to-Vehicle Communication.

V2X Vehicle-to-Everything Communication.

VANET Vehicular Ad Hoc Network.

WLAN Wireless Local Area Network.

Chapter 1

Introduction

> ”Information is the oil of the 21st century, and analytics is the combustion engine.”
>
> *Peter Sondergaard*

The ubiquity of mobile phones today is producing ever-increasing amounts of data, providing invaluable information that can be used to study many aspects of our everyday life. According to estimates, there were approximately 7.5 billion mobile subscriptions in 2016, with approximately 7 billion people living within areas with mobile network coverage [1]. Of these 7.5 billion subscribers, 4.5 billion are estimated to be unique users. Figure 1.1 shows the trend of subscribers, world population and population in coverage range of mobile networks. According to the 2017 Cisco Visual Networking Index [2], mobile data traffic has grown 18-fold over the past five years, and with mobile network coverage approaching 100% of the world population, there are increasing efforts to draw benefits from the data that the networks generate.

As the data is generated by large parts of the population, it is representative of many phenomena in our daily lives, and mobile network operators (MNOs) can create new revenue streams and aid developing countries [3]. In this vein, MNOs have recently launched research challenges, providing access to large data sets in an effort to extract knowledge therefrom for the benefit of various application domains [4, 5]. An extensive survey on the results in different fields – many of which are based on these research challenges – is provided by Naboulsi et al. [6], showing the vast scope of analyses that mobile data can support, ranging from social studies (e.g. demographics, land use and epidemiology) to

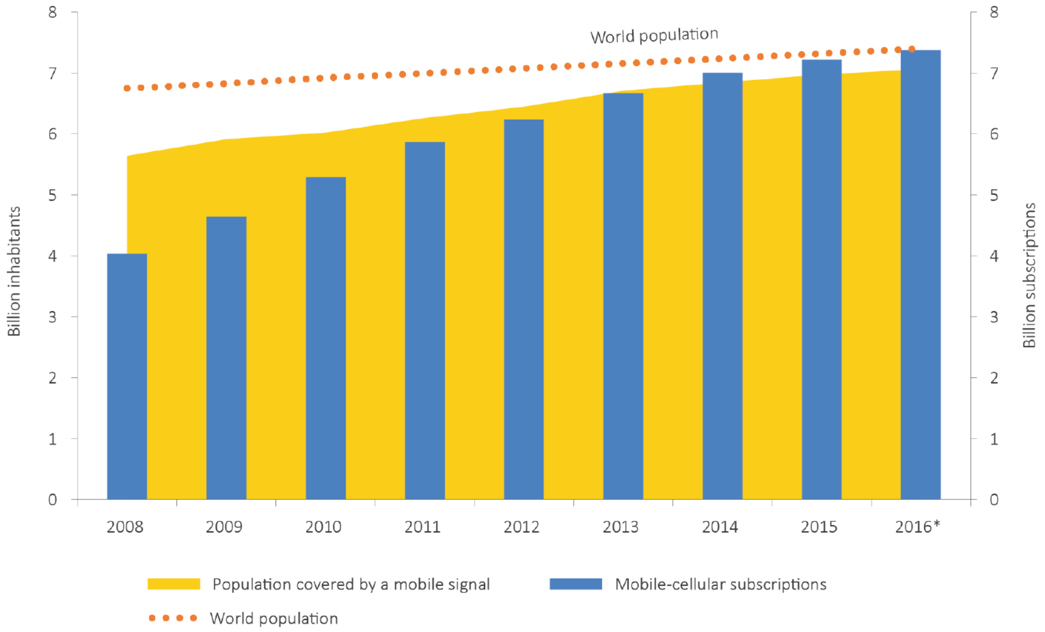


Figure 1.1: Evolution of global access to mobile network connectivity [1]

mobility and performance of the mobile network itself. Examples include techniques to facilitate urban planning [7], improve transportation networks [8, 9] and provide insights into the epidemiology of diseases [10].

In another survey on the analysis of mobile network data sets, Calabrese et al. [11] identify numerous topics of interest and open challenges. A common topic of interest between both surveys is the modelling of human mobility patterns, such as the use of mobile network data as a complementary source for estimating dynamic road traffic conditions. As a particular open challenge, they mention the characterization of the interplay between mobile networks and the actual mobility of users, while taking privacy and data anonymity considerations into account. Naboulsi et al. confirm that this is especially challenging for urban mobility, as it is much more heterogeneous in infrastructure and user behaviour than highway networks [6].

A major motivation for estimating urban mobility from mobile data is that transportation sensing infrastructure is costly in terms of installation and maintenance, has limited coverage and is intrusive. Thus, it is helpful to consider using mobile networks as a distributed sensor network for gathering data on mobility. In a mobile network, all participants generate data, i.e. stationary and mobile users alike. From a transportation perspective, mobile network data can then be referred to as *exogenic* data, as the data encompasses not just the transportation network agents but the entire userbase around it. This is in contrast to *endogenic data* generated purely within a transportation system, e.g. via Vehicle-to-Infrastructure Communication (V2I), and providing signals from certain types of users, such as cars or pedestrians exclusively. The main difference between the two approaches thus lies in the way the data is collected: exogenic data is passively generated, while endogenic data is purposefully, actively generated by a specific subset of users.

As the discussion on the most appropriate communication technology for vehicular communication is still ongoing [12], it makes sense to look at the passively collected exogenic data from mobile networks instead. In this context, and in order to leverage mobile networks and the data generated by its users *today*, it is necessary to study various types of mobile network data and their utility in estimating various transportation metrics. The outcome of this is the following research question:

1.1 Research Question

How do mobile and transportation network behaviours correlate and how can we leverage their interplay for transportation applications?

We want to explore how mobile network data can be used for various estimation and modelling tasks in transportation. Essentially, by looking at mobile networks as distributed traffic sensors, we want to show that they can serve as a complement to the existing, traditional transportation data sources. The aim is to use mobile network data directly, in contrast to data from connected car communication, as the latter is still sparse due to low equipment rates and the ongoing technology dispute. Thus, in the currently ongoing transition phase, exploiting exogenic communication data can provide significant knowledge.

Two main aspects of this research question must be considered.

On one hand, when characterizing the links between the behaviours of transportation networks in relation to that of the mobile network, it is important to measure the impact of stationary users. The question is whether data concerning the full mobile network userbase can be useful for describing only the mobile users, or whether the bias introduced by stationary users is too high. For certain types of data, e.g. regarding connection handovers, stationary users may be negligible in comparison to moving users, but the error must be quantified and is likely much higher for other types of mobile data.

On the other hand, it is necessary to investigate which estimation techniques are possible using different types of mobile network data. These types of data can be extracted from the billing system, e.g. CDR, and the core Radio Access Network (RAN), i.e. *signaling data*, and have different uses when it comes to transportation applications. The main applications with respect to private transportation supply are travel time and traffic state estimation. While these metrics are related, travel time estimation is usually concerned only with predicting delays and finding the fastest route, while traffic state estimation refers to the estimation of flow, density and velocity on a road (sub-)network. These metrics are useful for traffic engineering and control, and are key inputs for any Intelligent Transportation System (ITS). Demand estimation, on the other hand, typically involves all available modes of transportation, and is most frequently delivered as a set of dynamic O-D matrices that contain the need for transportation between different zones. Demand estimation is relevant for transportation optimization and planning, and is the key component for realistic simulation of transportation systems as well.

We want to explore ways of estimating supply- and demand-related metrics in a cost-neutral and privacy-friendly way. Mobile networks, when used as distributed traffic sensors, are one way of achieving this aim. Thus, we will present mobile network data-based models, in order to show how the main transportation metrics can be estimated. Mobile networks can then be used either alone or jointly with additional data sources to provide accurate estimates of transportation metrics, in rural and urban areas alike.

This research question is of particular interest for applications in areas with little transportation information infrastructure, e.g. in remote areas or in developing countries. For example, according to the 2017 Cisco Visual Networking Index [2], mobile data traffic has almost doubled in 2016 in the Middle East and Africa, a cumulative annual growth rate of 65% is forecast for the coming five years. Combined with the current, high coverage rate, as shown in Figure 1.1, we can expect mobile data to become increasingly valuable for knowledge discovery. The mobile network operators' data challenges and research efforts show that they share this view [4,5]. Thus, the research question has commercial potential, as there is real value to be extracted from mobile data.

In general, exploiting mobile network data is useful for transportation agencies, as it comes at low cost and with high spatial coverage, and it can provide additional insight.

1.2 Methodology

We evaluate the utility of mobile data for estimating transportation metrics of supply and demand in both simulated or synthetic, and real data.

First, in Chapter 3, we introduce a simulation package that was created for the purpose of performing studies for this dissertation, in particular for the traffic state estimation in Chapter 6.

Before using real data we performed studies on synthetic and/or simulated data for the main transportation metrics that we considered, i.e. traffic state, travel times and demand. Our studies on traffic states and travel times concern road traffic (private transportation), while we provide methods for demand estimation that can consider public transportation as well.

The types of data that we base our studies on are the following:

- Call Detail Records (CDR) – Billing Data
- Aggregated Handover Counts – Signalling Data
- Floating Car Data (FCD) – Road Traffic Ground Truth

Initially, we worked on CDR data, as it is the most widely adopted type of data in mobile data analysis. We based our first study on CDRs from the 2015 D4D challenge [4], then proceeded to generate synthetic CDT data from Floating-Car Data (FCD) and CDR data for the estimation of travel times and activity duration in demand estimation. The rationale behind this choice is that CDT offers a direct way of modeling the mobility behavior of mobile network users, conditioned on their current and/or previous or next locations.

For road traffic state estimation, we decided to use concepts from traffic flow theory, which rely on aggregated flows. With privacy concerns in mind, we opted for aggregated handover data, which was kindly provided by a Luxembourgish Mobile Network Operator (MNO). The idea behind this choice was to abstract individual user movements and instead consider handovers inside the network to approximate flows.

To summarize, in all our studies, we built models that are privacy-neutral, as they are based on aggregated data, and we synthesized the contained information into fitted model parameters. We demonstrate the utility of the proposed mobile network data-based models in predicting different transportation supply and demand metrics. In the case of traffic state estimation, we also compare real data and simulation results.

1.3 Contributions

The first contribution consists of a simulation scenario of Luxembourg City with LTE mobile network infrastructure, which allows jointly simulating cars' mobility and their connectivity to the LTE network. This contribution is presented in Chapter 3, and was published in [Derrmann et al., 2016a]. The simulation package is freely available online [1].

The second contribution consists of a study of the adequacy of mobile phone cell dwell times for travel time estimation. Based on synthetic data generated from FCD, we show that distributions of cell dwell times are a promising, privacy-friendly predictor for travel times in an urban setting. This contribution is presented in Chapter 4, and was published in [Derrmann et al., 2016b].

In the third contribution, we show that CDR data can serve as a valuable input for travel demand estimation, in particular the estimation of activity durations. This contribution models the temporal aspect of mobility, while the spatial aspect was handled by Di Donna et al. [13] within the joint MAMBA project framework. This contribution is presented in Chapter 5, where we additionally show how handovers can be used in activity-based demand estimation. This work was published in [Cantelmo et al., 2017].

The final, *main contribution* of this work is the traffic state estimation model we present in Chapter 6. We introduce a methodology for estimating vehicular density and flows in analogy to the Macroscopic Fundamental Diagram, using mobile network handovers as input data. We show that the presented model works both in simulated- and real-data settings, and compare the results from both worlds. This contribution was published in [Derrmann et al., 2017b, Derrmann et al., 2017a, Derrmann et al., 2017c], and the results were consolidated in a journal article submission that is currently undergoing review.

1.4 Structure

This dissertation is organized as follows. In Chapter 2 we present research related to the topic of using Information and Communication Technology (ICT) data analytics in ITS, introducing the required transportation models and show how different communication networks can be used for fitting transportation models.

In Chapter 3, we present the LuST-LTE simulation package, which produces realistic mobile and road network traces for Luxembourg City, and we perform a preliminary

[1] https://github.com/tderrmann/LuSTLTE/tree/LustLTEmod

validation of its behaviour.

Our first study in Chapter 4 shows how travel time estimation can be performed with mobile network data, yielding some limited, but promising results based on synthetic mobile network signaling data.

In Chapter 5, we show how mobile network data can support demand estimation techniques, and how it can be used for generating synthetic population data.

Chapter 6 presents a novel model linking traffic flow theory and mobile network data, which allows estimating traffic states on a macroscopic scale from mobile network signalling data.

We conclude our work in Chapter 7, summarizing the findings and giving perspectives for future work in this field.

Chapter 2

Intelligent Transportation Systems and Communication Technologies

Intelligent Transportation System (ITS) are defined as transportation systems that rely on Information and Communication Technology (ICT) to leverage information exchange between participants. The purpose of an ITS is the optimization of the performance of a transportation system with respect to specific target attributes such as average travel time or operational cost. This goal is achieved by sourcing sensor input data, evaluating supply and demand models of the transportation network, and using actuators to influence the behaviour of the network and enforce control policies. In the following sections, we provide a non-exhaustive introduction of these models and the data sources which they are based on, as well as the role of communication networks in the improvement of today's transportation networks.

2.1 Transportation Models

Generally speaking, the models used for transportation systems can be grouped into three distinct groups:

- **Supply** models describe the available capacity – static or dynamic – of a transportation mode or (sub-)network in terms of passengers or vehicles with respect to different factors influencing the state of the network.
- **Demand** models describe the need for transportation in terms of passengers, freight volume or weight, depending on different environmental and spatio-temporal factors.

- **Assignment** models allocate demand to the available supply resources.

In this dissertation, we concentrate on supply and demand model parameters from mobile data, which we now introduce in more detail. For a more in-depth insight into this topic, we suggest the books by Sheffi [14], Ortuzar and Willemsen [15] and Cascetta [16].

2.1.1 Demand Models

Demand models allow synthetic representations of the need for transportation in a given network to be created. The two main categories of demand estimation models are the *trip-based* and *activity-based* variations.

Trip-based demand models, the most established of which is the four-step model [17], consider trips independently from each other. The four-step model is defined by the steps of trip generation, distribution, mode choice and network assignment. Trip generation denotes the production and attraction of trips by zones. Trip distribution then allocates trips between pairs of zones, e.g. using the gravity model, yielding an Origin-Destination (O-D) matrix. Finally, mode choice and network assignment distribute the identified demand onto different transportation modes and route alternatives.

Recently, the focus of research has shifted towards behavioural models, i.e. *activity-based demand models*. They reproduce demand through the mobility needs, as defined by activity sequences. Through the chain of individuals' locations and the duration of their activities in those locations, the aggregate demand can be extracted. These models are most helpful in estimating mode choice, as they give a detailed image of user activity chains.

In [18], Toledo et al. present a recent overview of demand estimation techniques, both for static and dynamic (time-varying) estimation. They give an overview of the input data and optimization methods used in both congested and uncongested networks. The two cases are treated differently in demand modelling. In the uncongested case, a model of the transportation network provides link travel time estimates, which are then used for calibrating the demand estimation based on link counts and survey data. In the congested case, link speeds need to be considered for calibration, and as there is a bidirectional dependence between the O-D matrices and the traffic assignment, this case is more complex to estimate. One problematic aspect in demand estimation is the need for a seed O-D matrix providing an informative starting point that leads to a sensible optimization outcome. We will discuss this aspect in this dissertation in Section 5. Further, it is difficult to compare the performance of different demand estimation techniques in the absence of sufficient observational data. In this context, Antoniou et al. provide an extensive survey of demand estimation methods, and propose a framework for comparing different O-D estimation schemes [19], avoiding the problem that O-D flows are difficult to observe in urban networks. In summary, the ground truth is difficult to establish and different demand estimation schemes are difficult to compare.

2.1.2 Supply Models

Supply models describe the condition of the transportation infrastructure with respect to the demand to which it is subjected. When considering private transportation, these models typically stem from traffic flow theory, characterizing the traffic flow, density and velocity of a given segment or sub-network, which behave in a highly non-linear way. For public transportation, they typically describe the available transportation supply in a given mode or on a specific corridor, described e.g. by transit schedules or real-time passenger data.

Traffic State

Traffic State Estimation denotes the characterization of flow, density and velocity and their relationship on a road network partition or segment.

Depending on the type of road network to be evaluated, different models are employed for state estimation. For highways, the Lighthill-Whitham-Richards model (LWR), and other, related PDE models are primarily used to estimate traffic states between measurement points [20]. Urban road segments and networks are typically described by their flow-density relationship, i.e. the Fundamental Diagram of Traffic Flow (FD), going back to seminal studies by Greenshields [21]. For subnetworks, traffic measurements of a subset of the contained individual segments can be aggregated to form the Macroscopic Fundamental Diagram (MFD) [22] – sometimes also called Network Fundamental Diagram (NFD) [23] – of the region. They exhibit lower variance than the individual detectors, as the effects of local traffic phenomena are averaged. While they only emerge under certain conditions [24], MFDs currently are one of the main focal points of traffic flow theory research, since they are powerful tools for traffic forecasting and control [25, 26].

Travel Time

Travel Time Estimation considers estimating the duration – typically with quantified uncertainty – of travel between two locations in the network. Most commonly, in Advanced Traveller Information System (ATIS), travel time indications are used to inform users about the prospective duration of their planned trips. Routes are typically computed using graph-based data structures, where link weights represent segment travel times, and in which routing can be performed using established algorithms such as A* and Dijkstra [27].

In a *data-driven* approach, the predictions of link travel speeds can be made with various supervised learning models, ranging from regression methods such as ARIMA [28], to Kalman filters [29], neural networks (e.g. Long Short-Term Memory (LSTM) models [30]), and boosted regression tree methods [31]. The two latter methods can automatically identify relevant correlations between different traffic measurement points if the classifiers are provided with short-term historical observations from nearby sensor locations. In data-rich environments, i.e. transportation networks with high sensing infrastructure coverage, this is a practical alternative to model-driven forecasting.

In cases where observations are sparse, travel time estimation can be powered by *traffic flow theory*, producing better predictions than purely data-driven approaches. In highway travel time prediction, the Lighthill-Whitham-Richards partial differential equation can be used. Work et al. proposed a purely velocity-driven approach, by using the Greenshields fundamental diagram [21], enabling GPS to be a sufficient data source for velocity and thus travel time estimation on highways [20]. For arterial urban networks, Hofleitner et al. have proposed the use of particle filters for inferring the most likely state of intersections and subnetworks for forecasting purposes, also relying on fundamental diagrams and queuing theory [32].

Travel time estimation is performed for ATIS applications, but also in control-based transportation optimization, which we discuss in more detail in the following section.

2.2 Intelligent Transportation System Applications

The transportation models described above are used in the improvement of transportation in two main categories, i.e. transportation planning and on-the-fly optimization of operation through control policies.

2.2.1 Transportation Planning

Transportation planning describes the process of designing transportation networks and its service characteristics, such as schedules, frequency and stop locations. Depending on the temporal scope of the planning measures, they are commonly referred to as short-, medium- or long-term. While short- and medium-term planning involves the optimization and extension of the existing traffic infrastructure, long-term planning involves new infrastructure and potentially new technology, e.g. autonomous vehicles. Transportation models enable informed decision-making in this phase. Here, we provide some examples of aspects of transportation planning that can be supported by transportation models:

- *Infrastructure planning and multimodality:*

  When using a sufficiently realistic demand model of a transportation network, the impact of modifying different parameters of that network can be evaluated using micro- and macroscopic simulation, using tools such as SUMO [33] and the commercial software PTV VISUM, but also agent-based tools like TRANSIMS [34] and MATSIM [35]. In this case, the increase of transportation utility can be quantified across the entire population of interest for different target cost functions. This allows multimodal transport to be assessed to measure the impact of modifications of service parameters. Typically, agent-based simulation is performed, allowing individual itineraries to be analyzed, regarding e.g. inter-mode waiting times, walking distance to nearby stops and adequacy of schedules with respect to the demand. By using different cost functions, various consequences of modifying transit supply can then be estimated, such as economic [36] and ecological impact [37]. In the earlier stages, i.e. during the design of a network, models can be used to support intelligent decision-making, e.g. travel time models, allowing a faster exploration of the network design space through search algorithms [38].

- *Evaluation of carpooling and ride-sharing schemes:*

  Today's concept of the "sharing economy" is also impacting mobility. Ride-sharing and carpooling as alternatives or complements to existing transportation modes are becoming commonplace. Agent-based simulation with realistic demand models allows the assessment of the interaction of carpooling with other transportation modes [39]. Carpooling also benefits from supply models, as this enables optimized routing and thus better allocation of passengers and routes to drivers [40].

2.2.2 Operational Control

Control algorithms come into play when existing networks are to be optimized. The optimization algorithms react to the current state of the network and aim to redistribute the demand or re-organize supply such that the overall utility of the transportation network increases. Since, for many of these control algorithms, interaction between vehicles and a central entity is required, both mobility and communication need to be modelled. Thus, simulation tools like Veins [41] or LiMoSim [42] are used to evaluate the efficiency and compare the quality of different control objectives and policies.

- *Gating and Traffic Light Control:*

  Supply models – in particular of traffic state – enable the optimization of traffic by controlling the flows between zones. This technique of managing critical in- and outflow segments is commonly referred to as *gating* [25]. Another method of mitigating traffic congestion through redistribution is dynamic traffic light control at intersections, reducing queuing time and strategically prioritizing specific flows [26]. These methods require a precise knowledge of the traffic states in the controlled and surrounding areas, rendered possible by ITS and ICT.

- *Centralized Routing:*

  Through extensive sensor coverage, as provided e.g. by disseminated Floating-Car Data (FCD) [43], a central entity can coordinate traffic by providing optimized route recommendations aiming to globally optimize metrics such as travel time. The optimization algorithms often stem from the domain of game theory, as the agents in the system want to maximize their utility and the coordinator wants to achieve the system equilibrium point [44]. Incentives can then reward users for taking route alternatives that are sub-optimal from their individual point of view, but benefit the system's overall performance.

- *Variable Speed Limit:*

  Variable Speed Limit (VSL) is a valuable control tool for highway congestion mitigation, and reducing the overall accident risk [45]. Central coordination allows avoiding the onset phases of traffic jams by reducing inflow speeds, in turn reducing the effects of over-braking and phantom traffic jam occurrence.

  Traffic jam shockwaves can also be mitigated using a Cooperative Advanced Driver Assistance System (CADAS) leveraging Vehicle-to-Vehicle Communication (V2V) to redistribute traffic density longitudinally [46]. This has the advantage of not requiring a central supervision entity. Instead, local communication between vehicles, i.e. V2V, is sufficient.

- *Lane Reversal:*

  Lane reversal, i.e. the dynamic assignment of direction to a road segment is a promising concept in the coming availbility of autonomous vehicles. With autonomous vehicles and Vehicle-to-Infrastructure Communication (V2I), the risk of drivers not noticing the lane reversal and causing accidents is eliminated, and the benefits of adapting road network infrastructure to the demand can be reaped [47]. In order to enable lane reversal and the re-allocation of roads or areas to different directions or transportation modes, an exact knowledge of the traffic state is required. This can only be achieved through ITS and ICT and represents the highest level of control of a transportation system, adjusting both supply and assignment with respect to demand.

- *Public Transportation Control:*

  Public transport can also be targeted by control algorithms to optimize quality of service. For example, in [48], Cortés et al. show how predictive modelling of the number of passengers boarding and exiting buses can be leveraged to control bus station skipping and holding patterns, thus minimizing waiting times at stops and bus bunching phenomena. In the same vein, Dessouky et al. show in a simulation-based study how different types of information can be used for centralized coordination of buses [49]. They compare how bus tracking, passenger counting and ICT can enable improved transportation service over purely local observations of buses, and demonstrate that control is a necessity for optimizing traveller experience.

2.2.3 Convential Traffic Data Sources

For the different transportation models that we mentioned above, estimation is traditionally performed using dedicated sensing systems.

- *Loop detectors:*

  Induction loops embedded into roads can measure vehicular flows and the fraction of time that they are occupied. This allows traffic states for individual links to be computed, as density can be derived from occupancy. Loop detectors are the most common traffic sensing infrastructure, as they are also used for traffic light control systems.

- *Traffic cameras:*

  By means of image processing, traffic cameras can be an additional data source for approximating velocity and traffic density. Cameras are primarily used for safety-related applications, in particular on highways and in tunnels. Techniques for traffic estimation based on computer vision have been available for a long time, and with the increasing availability of consumer-grade GPUs capable of evaluating these models in near real-time, traffic camera data has become increasingly attractive for traffic state estimation [50].

- *Floating Car and other GPS data:*

  FCD provides vehicles' locations and movement vectors, allowing traffic engineers to reconstruct their trajectories and give estimates of link travel speeds. However, unless data from a significant percentage of vehicles is available, it is difficult to accurately estimate traffic density from FCD alone, and other data sources or V2V are used [43].

  With today's mobile phones being used for navigation and fleet management services, *Floating Phone Data* has recently become more valuable. It is the main input source for traffic information in large ATIS deployments as provided by Google Live Traffic or Waze.

- *Terrestrial Trunked Radio and other radio technologies:*

  Terrestrial Trunked Radio (TETRA) is a robust, long-range radio technology. It is mostly used for safety-specific applications, but commercial solutions for bus telemetry also exist. This allows transportation agencies to let travellers know about expected arrival times of buses at stops, but can also serve as input for controlling holding times and distance between successive buses. Data Radio Channel (DARC) is another technology used for bus telemetry, and uses FM radio to transmit digital data. Depending on the specific transportation solution, buses communicate directly with nearby panels at stops, or communicate with a central entity. In ATIS, bus telemetry data is also integrated as real-time information, e.g. in the General Transit Feed Specification: Real-Time (GTFS-RT) format, to provide travellers with current bus locations and average delays by time of day.

- *Static Bluetooth probes:*

  By monitoring the Bluetooth discovery process from a stationary sensor, a Bluetooth probe can serve in the same way as a loop detector, measuring the number of Bluetooth-enabled mobile devices passing by [51, 52]. In this context, Friesen et al. present a survey on applications of Bluetooth in ITS [53], showing how a typical Bluetooth probe can be designed, and presenting the results of different studies, primarily in travel time estimation. Note that the mentioned studies are based on stationary Bluetooth probes. In the following section, we discuss research on mobile Bluetooth probing.

- *Smart cards and RFID-based passenger data:* Pelletier et al. [54] provide an extensive literature review of the usage of smart cards in public transportation. The data that these systems generate is valuable for different purposes in planning and control, in particular demand modelling, since O-D matrices and mode choice can partially be estimated from them [55]. Further, network performance can be assessed using smart cards as a data source, as well as travel behaviour, i.e. trip purposes and sequences of transit users [56, 57].

2.3 Communication Technologies for Intelligent Transportation Systems

There have been multiple decades of transportation research to leverage the data sources mentioned above to estimate transportation parameters and fit supply, demand and assignment models. In order to make the most of the information provided by these models, operational optimization often requires (partial) connectivity between the agents and the transportation agencies, allowing intervention in the behaviour of the transportation systems.

Intelligent Transportation Systems serve to shape supply and demand so as to guarantee better quality of service, or optimize other goals, e.g. ecological targets. Examples of optimization objectives are the reduction of overall travel times in the system, the minimization of waiting times at stops, calming of traffic and the limitation of pollution in cities.

Transportation agencies can intervene both on the demand and supply sides. Transportation demand can be influenced through a set of policies commonly known as Transportation Demand Management (TDM). Common TDM measures are incentive-based approaches such as congestion pricing, the modification of transit supply for temporal demand distribution and the support and incentivization of carpooling, ride-sharing and electric mobility.

Through central orchestration or peer-to-peer coordination, it is possible to support these measures and enable better decision-making. In order to enable real-time intervention of transportation agencies, there is a need for real-time information and interaction between agents in the transportation system. Cooperative Intelligent Transportation Systems (C-ITS) designates the concept of coordinating transportation by the means of ICT, in particular using V2V and V2I. There are various optimization objectives that can be addressed through cooperative ITS, ranging from ecological footprint reduction (CO_2 emissions) [58] to collaborative or centrally-orchestrated routing of traffic striving for a global traffic equilibrium [44].

Through the increased availability and use of communication technologies, the metadata generated from them has become a valuable source of insight for various applications. In the context of ITS, it is necessary to distinguish between two types of data. On one hand, there is the data created within the transportation system by its agents, i.e. *endogenous data*. On the other hand, there is data that is continuously generated by surrounding devices and infrastructure, i.e. *exogenous data* from outside the ITS, which can serve as an additional information medium or distributed sensing mechanism.

Due to the currently ongoing dispute about which technology will be used for V2X communication in the future, it will take 5-10 years to be widely adopted in vehicles [12]. Consequently, in the meantime, it is sensible to look at data from existing communication technologies and infrastructures and see how they can support the estimation of transportation metrics. For transportation engineers, the main advantage of passively-collected communication network data lies in the penetration rate of connected devices. Unlike traditional methods such as travel surveys, they cover a large proportion of the population at the potential cost of being biased with respect to certain transportation modes and stationary or low-mobility users. Additionally, they are mostly cost-neutral, as there is no need for additional infrastructure to be put in place. When combined with traditional data sources, communication network data can lead to a wide range of applications, some of which we present in the following paragraphs.

Bluetooth

Bluetooth, in its different versions Bluetooth Classic (BC) and Bluetooth Low Energy (BLE), is a pervasive technology and increasingly popular. Today, there are over 2.5 billion Bluetooth-enabled mobile phones [59]. The main application of Bluetooth for transportation sensing is by using data generated during the *device discovery* process, through which devices detect each other and the services they offer, e.g. audio or data streaming. This discovery process makes it possible to count the number of nearby devices through the number of distinct discovered Medium Access Control (MAC) addresses, and to approximate their distance or relative speed using the Received Signal Strength Indicator (RSSI).

Friesen et al. present a survey on applications of Bluetooth in ITS [53]. From a mobile device perspective, Bluetooth discovery data can be used for contextualization, i.e. to classify the type of environment that a vehicle or person is in, e.g. riding a bus or driving on a highway. Bronzi et al. showed the feasibility of identifying the road class by considering features extracted from BC and BLE discovery data [59].

Wireless LAN

Wireless Local Area Network (WLAN) technology is built into most mobile devices available today, and the Access Point (AP) density in urban areas is so high that the Service Set Identifier (SSID) names are used to improve GPS localization resolution. It is therefore unsurprising that transportation researchers are looking at WLAN as a potential source of information, in particular for quantifying pedestrian movements. Similar as with Bluetooth, the probe request mechanism for discovering nearby APs is of interest for passive sensing.

By detecting WLAN probe requests, it is possible to extract knowledge concerning pedestrian flows in the vicinity. This is of particular interest for companies operating multiple APs, who are able to track users' locations and their signal quality. With respect to ITS, WLAN probing can be leveraged in a similar manner as Bluetooth data, e.g. for detecting pedestrian mobility behaviour, or waiting times at bus stops or train stations [60]. In this way, identifying the number of people at bus stops can help in establishing demand patterns for public transportation.

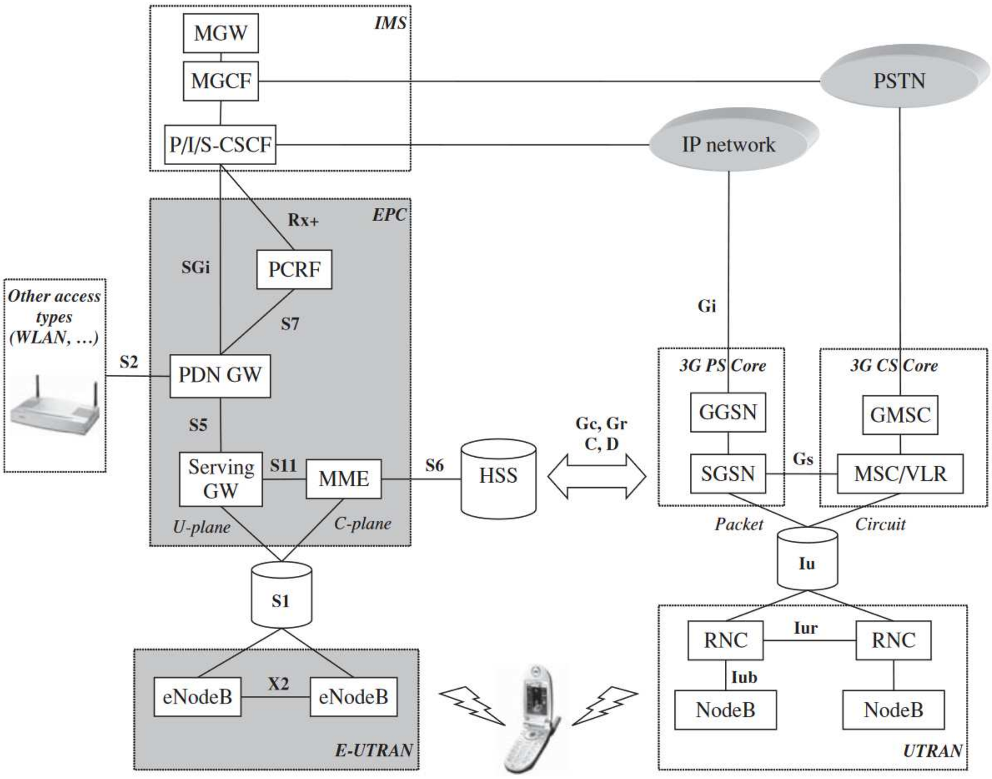


Figure 2.1: Simplified diagram of a UMTS and LTE network architecture [61]

Mobile Networks

Mobile phone networks, due to their pervasive coverage and hierarchical architecture, are an invaluable source of information regarding mobility. Radio access infrastructure is placed so as to guarantee the best possible quality of service. As the infrastructure must scale with commuters and mobility in general, base stations cover the main corridors of traffic and population and business centres.

Figure 2.1 shows the typical architecture of a mobile network with Universal Mobile Telecommunications System (UMTS) and Long-Term Evolution (LTE) Radio Access Network (RAN). The left side shows the LTE network, while the right shows the UMTS infrastructure. For the sake of conciseness, we limit our discussion to the network components relevant to the collection of data as used in this dissertation. For in-depth explanations of the other network components and the abbreviations used, we refer the reader to [61] and [6].

Caller ID	Callee ID	Timestamp	Cell ID	Duration/Volume	Type
123456	456789	1513344059	999	336.0	Call
234567	111144	1513344063	765	1.0	SMS

Table 2.1: Call Detail Record Example

Generally speaking, there are two categories of mobile network data that are used in research, namely *Billing Data* and *Signalling Data*. Billing data is typically collected from the Charging Gateway Function (CGF), which may be distributed within the SGSN or GGSN (Serving/Gateway GPRS Support Nodes) or Serving or Packet Data Network Gateways (S-GW and PDN-GW). The billing data usually comes in the form of Call Detail Records (CDR). Table 2.1 shows the typical fields of a Call Detail Record. Note that CDRs contain additional information, e.g. the call result and fault codes, which we omitted from this listing for readability. A single CDR entry corresponds to a (usually) billable operation performed on the network, thus individual user observations are usually quite sparse, but the aggregate statistics are very useful. The applications of CDR data analysis are extremely diverse, ranging from mobility modelling to social network analysis and population density estimation, as enumerated in the surveys by Naboulsi et al. [6] and Calabrese et al. [62].

Signalling Data is typically extracted from probes closer to the Radio Network Controller (RNC) or Mobility Management Entity (MME). The signaling data that we work with in this dissertation is handover data, which is collected using a probe on the S1-MME and Iub interfaces of MME and RNC for LTE and UMTS, respectively. Today's mobile network monitoring hardware typically supports a multitude of protocols and can thus work with input data from different Radio Access Technologies (RATs). In contrast to Call Detail Records, handover data provides multiple sightings of an actively-connected user device, yielding a clearer picture of user trajectories. Note that in this dissertation, we only work with aggregated data, to avoid privacy caveats, while preserving information on macroscopic movements of mobile network users. Table 2.2 shows an example of such aggregated handover data, where number of handovers is summed up over an observation period for all cell source-destination pairs in the network.

At the beginning of the chapters on travel time (Chapter 4), demand (Chapter 5) and traffic state modelling (Chapter 6), we will specifically explore the state of the art of using mobile network data to model each of these transportation metrics.

Source Cell ID	**Destination Cell ID**	**StartTime**	**EndTime**	**Count**
123	456	1513344000	1513347600	250
987	123	1513344000	1513347600	720
123	456	1513347600	1513351200	340

Table 2.2: Aggregated Handover Data Example

Privacy Challenges

Both in WLAN and Bluetooth, there are significant privacy risks because many people inadvertantly leave their devices discoverable or constantly probing. This makes the reconstruction of user trajectories a real threat, as research has recently proven that as few as four data points locating an individual can lead to their identification [63, 64]. In Vehicular Ad Hoc Network (VANET) applications, this problem is typically avoided through pseudonyms that are used as temporary identifiers, coupled to a digital signature to prove message authenticity, and regularly renewed [65]. Similarly, unprocessed Call Detail Records link the International Mobile Subscriber Identifier (IMSI) and/or International Mobile Equipment Identifier (IMEI) to user locations, and potentially connecting this to the identity of an individual through the extraction of their home and work locations.

In this dissertation, we address the privacy problems through adequate aggregation of data, removing information of individual users. We aggregate the data with the target application in mind. By fitting different kinds of models to the data, we abstract the data to a level where it is privacy-neutral, but the core information on the target transportation metrics remains intact. Note, however, that there are other applications in which trajectories are required, and that there are anonymization techniques for user trajectories [66], but that this is not in the scope of this dissertation.

Chapter 3

Co-Simulating the Mobile and Road Networks

In order to study the link between mobile and road networks, simulation studies are invaluable, as they can lend first insights into whether a certain model may work in the field, with real data. In this dissertation, we want to show that mobile network connectivity can be used as a basis to passively observe the road network dynamics and mobility. However, to the best of our knowledge, there exist only few solutions to simulate such vehicular applications, especially if interaction between the cellular and the road networks is required. For this reason, we opted to augment the existing LuST scenario of Luxembourg City [67] with Long-Term Evolution (LTE) connectivity. We will now provide an overview of previous and novel simulation tools and scenarios for heterogeneous vehicular connectivity, and then introduce the LuST-LTE project that was developed for the purpose of performing the required studies for this dissertation.

3.1 State of the art

At the time when we performed this study, no simulation packages were available that could jointly simulate road network mobility and LTE connectivity with handovers. However, VeinsLTE [68] provides an environment that allows LTE communication, which we extended by a simple handover mechanism, as described in the next section. Also, with Luxembourg City as a potential use case study in mind, we opted to use the LuST project [67] as a starting point.

Other approaches involve reproducing mobility traces within a network simulator, such as the KölnTrace project, that provides 24 hours of mobility in the city of Cologne alongside the matching mobile network base station locations [69]. This scenario can then be used to reproduce the connectivity of vehicles in the simulation.

While Veins and LIMoSim use OMNeT++ for simulating the wireless networking part, there are also projects such as iTETRIS that rely on ns-3 for simulating WAVE/DSRC and mobile networking [58, 70].

Since the work on the LuST-LTE project decribed in this chapter, additional work was performed by other researchers. Recently, the SimuLTE developers have integrated Veins into a recent version of their project.

As a light-weight alternative for some use cases, the LIMoSim project was recently presented [42]. LIMoSim simulates vehicular mobility directly within OMNeT++. This is a different approach from that of Veins, which realizes on the full microscopic simulation of traffic in SUMO.

3.2 The LuST-LTE Project

We want to provide researchers with a simulation package that enables simulating pervasive vehicular access to LTE, enabling the evaluation of V2V and V2I protocols (e.g. for traffic routing). The simulation scenario we propose relies on VeinsLTE [68]. VeinsLTE is an extension of Veins [41], a software framework connecting the microscopic road traffic simulator SUMO [33] to the network simulator OMNeT++ [71], and providing support for the 802.11p car-to-car communication protocol. VeinsLTE extends Veins by adding support for the SimuLTE library [72] in OMNeT++, enabling simulated vehicles to have LTE connectivity.

Simulation tools and frameworks therefore need to use recent, realistic and scalable scenarios, where the user has the ability to fully characterize the network topology. Indeed, the vast amount of research in ad-hoc, mesh, sensor and cellular networks has shown that the network topology has a strong effect on the network performance, its reliability and its adaptation capacities [73–76]. The network density has, for instance, a direct effect on local congestion and can be challenging for the medium access control layer. For this reason, the scenario that we base our package on is LuST [77], a 24-hour road traffic scenario of Luxembourg City featuring realistic traffic behavior. By adding handover support to SimuLTE, and integrating eNodeB positions to the Luxembourg scenario, we

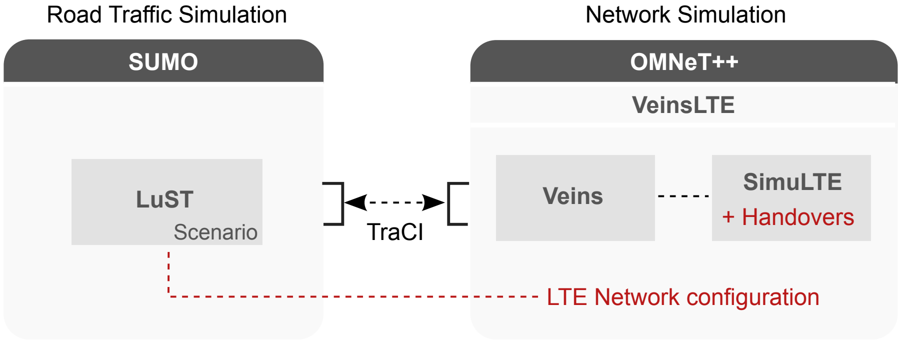


Figure 3.1: Overview of the LuST-LTE simulation package.

have created this package as a base for running studies on pervasive vehicular connectivity in a realistic setting.

3.2.1 Simulation Environment

The LuST-LTE simulation package bundles two components: road network and mobile network simulation. It comes with a pre-configured scenario of Luxembourg City, but is extensible to other scenarios as we will discuss later. Fig. 3.1 shows an overview of the package, and the interaction between the road and network simulations. LuST-LTE makes use of the VeinsLTE framework by Hagenauer et al. [41, 68] to connect both simulators. The road network simulation is performed by the microscopic road traffic simulator SUMO [33], configured to run the LuST scenario [77], while the mobile and vehicular network simulation uses the network simulator OMNeT++ [71], which runs SimuLTE [72]. We extended SimuLTE to include a simple handover mechanism, to allow pervasive vehicular access to the LTE network (e-UTRAN). We will now explain the different software components in more detail.

Road network: SUMO and LuST scenario

The LuST scenario by Codecà et al. [77] describes the road network, buildings and traffic demand of Luxembourg City and its highway ring, as input files for the Simulator of Urban Mobility (SUMO [33]). It is available as an open-source scenario[1] and covers a surface area of around $150km^2$. The scenario includes 38 bus lines, and the number of vehicles in a full 24-hour simulation run amounts to almost 300.000 vehicles.

[1] `http://vehicularlab.uni.lu/lust-scenario/`

Our goal is to enable researchers to evaluate heterogeneous vehicular applications, e.g. collaborative routing, in a ready-to-use package with a realistic scenario. In particular, the LuST scenario we base our package on is closer to the reality of city topologies than the frequently used grids. Moreover, other existing scenarios are generally limited in terms of duration and/or scalability. For instance, TAPASCologne[2] is one of the largest freely available traffic simulation scenario for SUMO. It models the city of Cologne (Köln, Germany) based on the OpenStreetMap cartography but is only limited to two hours of traffic (basic version). Uppoor et al. [78] studied this scenario to characterize vehicular RAN access using a Voronoi tessellation of base stations. By contrast, we aim to co-simulate the road and mobile networks to enable LTE-based vehicular applications.

Connecting road and network simulations: VeinsLTE

The Veins framework by Sommer et al. [41] acts as an intermediary framework to synchronize both simulators (SUMO and OMNeT++), and enables interaction between them. Additionally, Veins includes a model of IEEE 802.11p for inter-vehicular communication. This allows the simulation of diverse applications, allowing the communication network to influence the behavior and routing of vehicles.

LuST-LTE makes use of VeinsLTE, an extension of Veins by Hagenauer et al. [68], that extends Veins to also work with the OMNeT++ library SimuLTE. This enables cars to be LTE-equipped, and to communicate over E-UTRAN. The fact that VeinsLTE enables interaction between road and communication networks further enables applications such as intelligent traffic control using mobile network data. The next section describes the SimuLTE library and the modifications we included for LuST-LTE.

Mobile Network: SimuLTE

The LTE user plane is implemented by SimuLTE [72], which provides the LTE stack, i.e. physical layer, MAC, Radio Link Control (RLC), Packet Data Convergence Protocol (PDCP), and Radio Resource Control (RRC). Currently, it offers support for device-to-device communication, but no handover implementation yet. SimuLTE is integrated into the OMNeT++ INET framework, which is widely adopted. SimuLTE comes with a realistic channel model relying and different propagation model parameterizations depending on the type of eNodeB and environment. More precisely, SimuLTE uses the LTE path loss

[2] `http://sumo.dlr.de/wiki/Data/Scenarios/TAPASCologne`

models for line of sight and non line of sight situations, for different environments and cell sizes as defined in [79]. The propagation fading can be computed with different fading models, including the Jakes model for Rayleigh fading [80], which is an adequate model of urban environments and was used for the simulations in this study.

Extensions: Adding handovers and LTE infrastructure

In LuST-LTE, we included the actual eNodeB locations of a Luxembourg mobile network operator into the LuST scenario. That means that there is purely LTE infrastructure - no 2G or 3G - of a single mobile network operator. This yields a total of 38 eNodeBs for the scenario area.

Fig. 3.2 shows the distribution of eNodeBs along with the Luxembourg road network. A common model for coverage areas is the Voronoi tessellation, splitting the territory into areas that are closest to each eNodeB [6, 78]. Therefore, the Voronoi tessellation is also displayed, separating the terrain into polygons that are nearest to their enclosed eNodeB. Note that the topology of mobile phone network infrastructure is designed to optimize coverage and capacity with respect to population mobility. Thus, by their very design, this enables cellular networks as a distinguished data source for many studies. In Fig. 3.2, this is particularly visible by the cell density in the city center, compared to the more sparse coverage along the highway ring and the outskirts of the city. This means that we expect the signal strength should be higher and more stable in the city center, where there is a higher eNodeB density and vehicle speeds are lower. In section 3.2.2, we will take a closer look at those aspects.

In the current version of LuST-LTE, handovers are triggered based on the Signal-to-Noise-Ratio (SNR). In LTE, handovers and cell-reselection are based on RSRP and RSRQ (Reference Signal Received Power and Quality) measurements[3]. They define signal power and quality on a single reference signal, instead of all resource blocks as is the case for the SNR that we consider. However, Afroz et al. have shown in [81] that SNR can be a valid proportional substitute for RSRP. In future versions of the scenario, we might consider moving towards Reference Signal based handovers in favor of more realistic results and better comparability to real-world data. Also, in the current state, the scenario does not yet include information on the eNodeBs' transmission power levels. As this parameter is configurable in SimuLTE, we are planning to integrate it into the scenario once it is made

[3] http://laroccasolutions.com/training/78-rsrp-and-rsrq-measurement-in-lte/

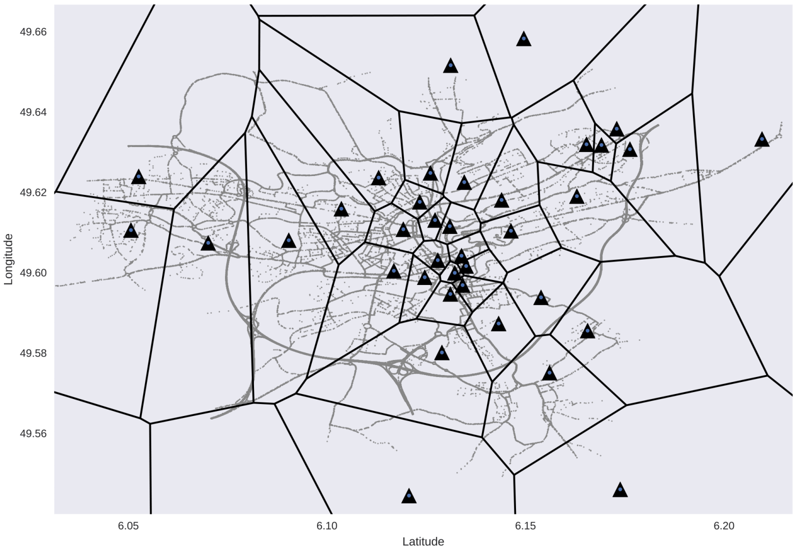


Figure 3.2: Map of eNodeB Locations and Voronoi Tessellation

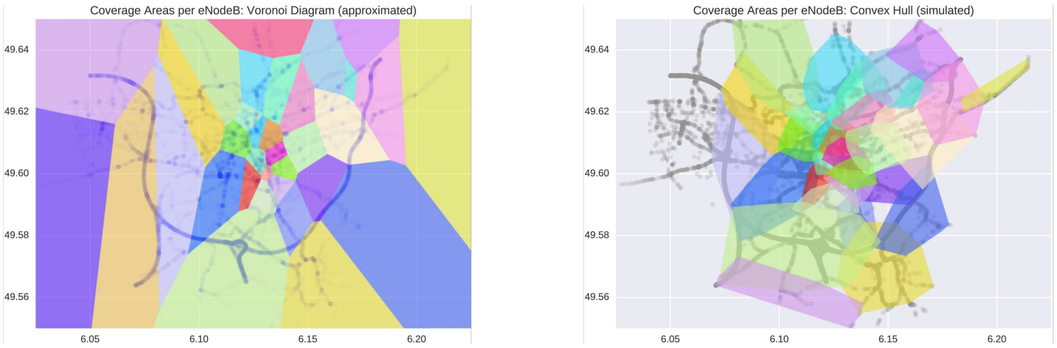
available to us by the mobile network operator.

Performance

The scenario's running time depends primarily on the LTE penetration rate, that can be adjusted in the scenario configuration files. For the results provided in this article, we used 5% vehicle equipment rate, as a good compromise between performance and result precision. This configuration yielded on average 0.3 simsecs/second on a current-generation 16-core, 32GB RAM server for the full mobility demand. While there is some potential of future optimization, this level of performance is sufficient as simulations are typically not dependent on real-time performance. The simulation performance can be improved by reducing the simulation scope in OMNeT++ (considering only vehicles within a specified geographical bounding box) or the number of vehicles in SUMO (by modifying the simulation start- and end times).

3.2.2 Evaluation

In this section, we will provide an evaluation of the simulation scenario with respect to cell sizes, signal strengths and network-wide statistics. The simulation was run without data exchange over LTE, only for the purpose of gathering signaling data. As baselines, we retrieved signal strength measurment data through the OpenCellID project, and also


Coverage Areas per eNodeB: Voronoi Diagram (approximated)
49.64
49.62
49.60
49.58
49.56
6.05
6.10
6.15
6.20
Coverage Areas per eNodeB: Convex Hull (simulated)
49.64
49.62
49.60
49.58
49.56
6.05
6.10
6.15
6.20

(a) Voronoi Tessellation

(b) Convex Hull

Figure 3.3: Coverage area comparison: Voronoi tessellation and simulated locations (represented by their convex hull)

refer to common approximation methods and previous results to confirm the plausibility of the simulation results.

Coverage Area

First, we look at the area covered by each eNodeB in the simulation, i.e. the area within which simulation vehicles were associated with each eNodeB. In Fig. 3.3, we compare the Voronoi tessellation with respect to antenna locations to the convex hulls of each eNodeB's coverage areas. Each color indicates a different eNodeB, and both plots follow the same color-map. The obvious similarity of both maps confirms that the handover mechanism is working as intended. Inversely, this also shows that the Voronoi tessellation approximates the simulated coverage area, confirming the adequacy of Voronoi tessellation for estimating base station coverage areas. Note that in the simulation, there is overlap between coverage areas, which is likely to increase if additional mechanisms such as power boosting/cell breathing are implemented.

We also looked at the number of unique UEs associated with each eNodeB over the course of a full simulation run. Looking at Fig. 3.4, we can see that there are two eNodeBs that are particulaly frequented. A dozen eNodeBs encounter more than 500 UEs (at 5% LTE equipment rate). The most important eNodeBs are located in the city center and in the south area of the ring, while the remaining two thirds of eNodeBs in the simulation have lower relative importance.

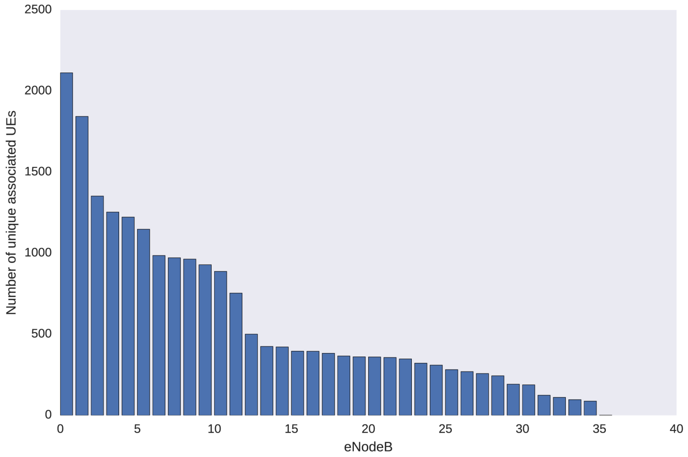


Figure 3.4: No. of unique associated UEs per simulated eNodeB at 5% equipment rate

Signal Strength - SNR

Fig. 3.5 shows color-coded signal strength and the contour of vehicular density levels in a joint representation. The density is evaluated over an entire day, and we can see the highest values on the ring and in the city center, in high coverage areas. In the south-west, we can observe low signal strength and the absence of a nearby eNodeB. However, there are nearby 3G base stations, so this could be due to the progressive equipment upgrade strategy of the mobile network operator.

Fig. 3.6 shows the correlation between the simulated signal-to-noise ratio (SNR) and real, measured RSRP as provided by the free OpenSignal.com Web-API [4]. We split the scenario territory into squares of 0.25 km^2, over which we averaged SNR and queried RSRP from the API, yielding around 100 measurement squares with available data from the required mobile network operator. As discussed by Afroz et al. in [81], SNR is proportional to RSRP in low network load situations, so this comparison is sensible for evaluating the signal strength generated in the scenario. The Pearson coefficient ρ amounts to 0.317, indicating moderate correlation between both metrics. The differences stem from the very low network load in our evaluation (unlike real data), along with the fact that the emission power of all eNodeBs in the simulation is identical, while it varies in practice (between eNodeBs, but also through cell breathing).

[4] www.opensignal.com

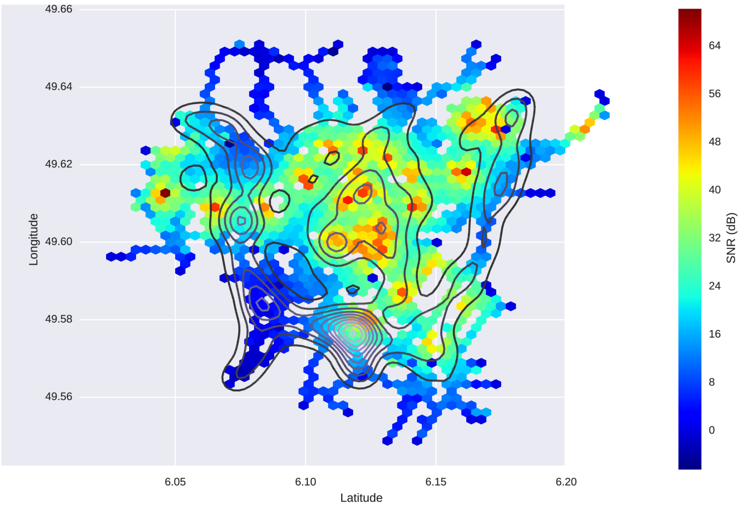


Figure 3.5: SNR vs. vehicular density kernel density estimation

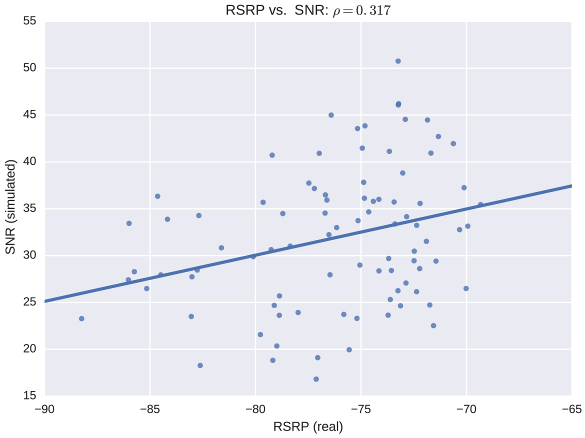


Figure 3.6: Simulated SNR vs. corresponding real RSRP

We also evaluated the signal propagation characteristics. Fig. 3.7 shows the joint distribution of SNR and the linear distance to the associated eNodeB. In the central joint distribution plot, we can observe the density following the propagation model's characteristic exponential curve. It shows how closer distance to the eNodeB results in higher SNR values. The variance is due to line of sight and non-line of sight situations due to buildings.

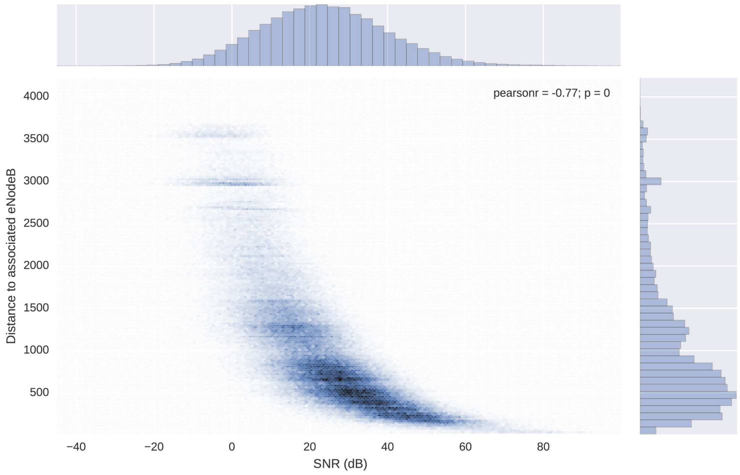


Figure 3.7: Simulated SNR vs. distance to associated eNodeB

Note that the marginal distribution of SNR (in the histogram at the top) appears to be following a normal distribution, with a mean at around 23dB. As this is close to the commonly used threshold for an excellent signal quality at 20dB[5], we can say that half of the values observed correspond to excellent signal conditions, and half of the datapoints correspond to weak to good connection signals.

6.3% of datapoints are located in the < 0dB area, which corresponds to edge-of-cell, low connection quality. The distance to the associated eNodeB follows a heavy-tailed distribution. This matches the intuition that the associated eNodeB is more likely to be nearby, with distances less than 1500km representing the majority (around 80%) of cases. The median value of 768m shows that half of the cases are likely urban, where eNodeBs are more densely distributed and thus more likely to be nearby.

Handovers and Dwell Time

We will now evaluate the simulated signalling data from the simulation run, i.e. handovers and cell dwell times. Cell dwell times are defined as the time a UE stays associated with an eNodeB before performing a handover. In the work leading up to this study, using floating-car data and using a Voronoi tessellation of 3G and 4G base stations, we have identified a proportionality between squared dwell times and handover counts. A visualization of this relationship between the resulting dwell times and handovers is displayed

[5] `http://laroccasolutions.com/training/164-rsrq-to-SINR/`

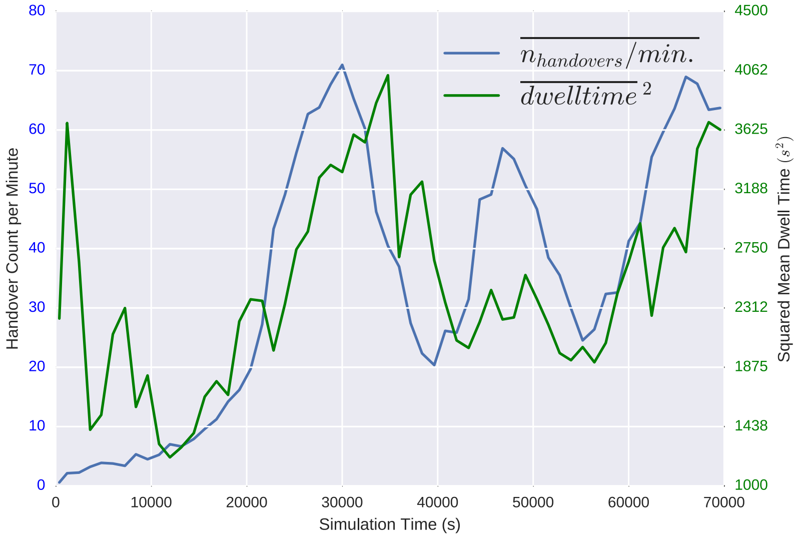


Figure 3.8: Squared mean dwell time vs. Number of handovers per minute

in Fig. 3.8. It shows squared mean dwell time with respect to the time of day (blue line) and the number of handovers per minute (green). Both lines are plotted with 20 minute intervals. Comparing them for the simulated data supports this with a Pearson correlation of $\rho = 0.58$. This correlation between the number of handovers ($\propto$ flow) and dwell time ($\propto$ speed^{-1}) indicates that there could be potential for the development of LTE signaling data-based road traffic estimation system.

Another interesting signaling metric is that of the number of unique eNodeBs that a vehicle associates with during a trip. In Fig.3.9, we show, for all trips of a day, the distribution of trip lengths and the number of distinct associated eNodeBs that the UE was connected to. The marginal distribution of trip lengths (top histogram) has high variance, showing the large variety in trips in the mobility. The marginal distribution of the number of connected eNodeBs appears to be bi-modal, which could be a characteristic separating trips that are pass by the city center (with high eNodeB density) from suburban and highway trips (with lower eNodeB density).

Looking at individual trips allows to further inspect the behavior of handovers in the simulation. Fig. 3.10 consists of two plots that describe a single user trip. The left-hand side plot shows the trajectory of a single vehicle, along with the locations where handovers performed and the associated eNodeBs. Note that the handovers are also visualized using colored lines that match between both plots. The right-side plot shows the SNR evolution

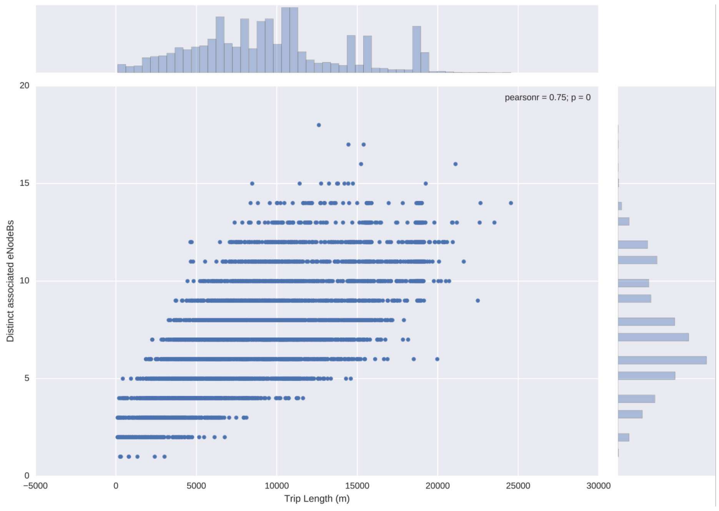


Figure 3.9: Number of distinct eNodeBs associated with during a trip vs. trip length

over time for a single vehicle trip (blue), along with the instants when handovers were triggered (vertical lines). Thus, it is possible to see where, when and at which signal strength the handovers occurred.

We can see that the signal levels along the trip are generally good and that handovers typically bring about an improvement in signal strength. However, we also observe quick successive reselection between 2 eNodeBs at around $t = 1700s$. On the map, we can see that this is due to the fact that the trajectory passes nearly centrally between both eNodeBs, causing a nearly immediate switch between them. This behavior can be attributed to the so-called *ping-pong* effect, where a User Equipment (UE) changes between 2 eNodeBs repeatedly in a single or nearby locations. It indicates that adjusting the signal quality hysteresis thresholds would yield more stable connectivity patterns by preventing UEs from changing between eNodeBs with immediate effect.

To summarize, we have verified that the handover mechanism works as intended, and that signal strengths are already usably realistic. There is some potential for improvement in the emission power of the urban eNodeBs, and we aim to include that information once it is made available to us by the mobile network operator. Also, we identified that a higher hysteresis factor might be considered for future revisions of the scenario, but there is a need for further real-world data (i.e. from the network operators).

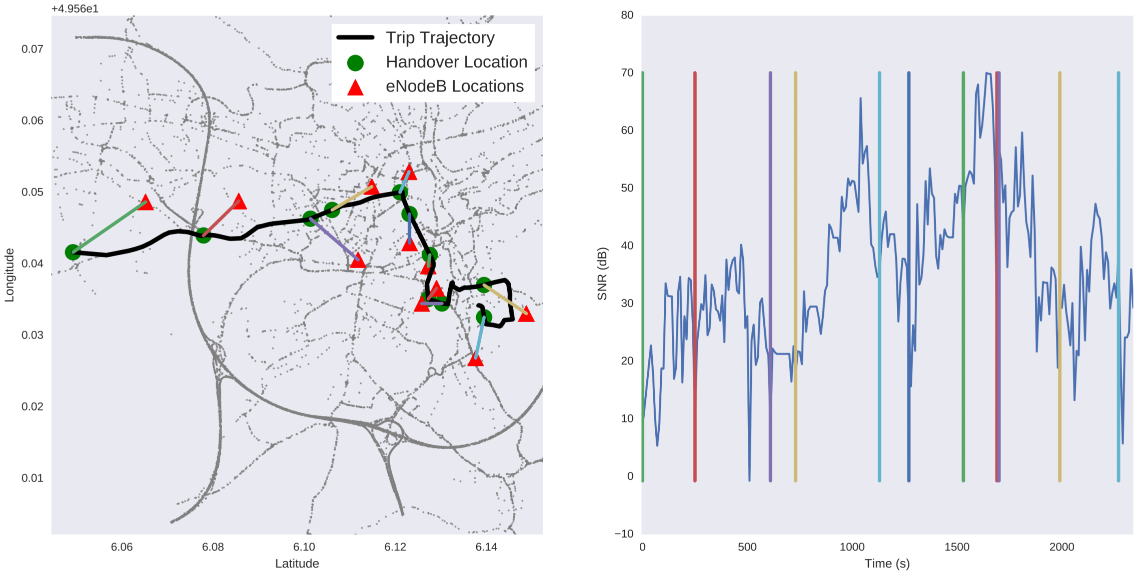


Figure 3.10: Example trip. Left: Trajectory and handovers; Right: SNR evolution (blue), handovers (vertical lines)

Chapter 4

Travel Time Modelling using Mobile Network Data

In this chapter, we investigate the potential of Cell Dwell Time (CDT) distributions in a mobile network to serve as travel time predictors of the underlying urban road network.

At the time when this study was performed, we did not have access to a real cell dwell time data set. Instead, we opted to synthesize such a data set from Floating-Car Data (FCD). Thus, the prediction errors presented in this chapter are lower than what could be achieved from real data, as they do not involve the step of filtering out local mobility (pedestrians and bikes) and stationary users. On the other hand, the penetration rate of users in the real data is much higher, leading to a higher statistical robustness, which may balance these restrictions out.

We also investigate the link between handovers and travel times in the synthetic dwell time data set, and suggest some visualization techniques for mobility based on dwell time data.

4.1 State of the Art

In the past, most studies on travel time estimation based their analyses on floating car data sets mainly comprising GPS data points [82], [83]. Over the last decade, research on cellular data analysis has gained significant popularity. The survey by Naboulsi et al. [84] provides a comprehensive overview of state-of-the-art results and methodologies. Supplementing this, the work by Valerio et al. [85] provides a literature review of cellular-based road monitoring system and proposes an extended monitoring framework for both circuit-switched (GSM) and packet-switched (3G and 4G) networks.

One method of capturing population dynamics on a city-wide scale was presented by Reades et al. [86]. Through a collaboration with Telecom Italia and MIT, the authors analyzed different mobility data sets of the city of Rome. The primary metric they used for their study was the Erlang, a measure of network bandwidth usage. Although this metric is highly aggregated and does not allow individual identities to be deduced, it provided interesting insights into the spatial and temporal dynamics of a city. Augmenting these methods with public and private transportation data yielded deep insights into population mobility [87].

Similarly, Trasarti et al. [88] analyzed anonymized CDRs provided by Orange France in order to detect connections between different locations that can be inferred by the spatial distribution of mobile phone activities. They introduced a new correlation metric (C-pattern) aiming to discover hidden logic of connections between different regions by analyzing frequently co-occurring changes in population densities. The resulting visual representation provides an aggregated view on the connection between different regions both on the urban and national level.

Another visualization method that can be applied to large data sets was introduced by Andrienko et al. [89]. This work tackles the problem of the processing of a large amount of location data points (e.g. GPS trajectories) in order to extract and visualize meaningful clusters. They propose a generic two steps approach involving a human analyst, which directs the work of the computer towards the discovery of meaningful clusters.

A more specific study on cellular data analysis to detect highway traffic congestion has been provided by Janecek et al. [90]. In this work, the authors propose an approach that combines several large-scale cellular data sets in order to detect and classify road congestion on a selected highway segment in Austria. To do this, they first rely on coarse-grained signalling data available from all idle terminals on the network to estimate travel time. This information is combined with fine-grained data provided by a subset of active terminals (e.g. performing a voice call) to localize and classify the congestion events. The results show that their approach outperforms other monitoring technologies in detecting road traffic congestion.

Schlaich et al. [91] show that demand for traffic models can be generated from Location Area Code (LAC) sequences. They conclude that this type of data is useful for trips with more than three LAC updates. With this filtered data set, it is possible to map LAC sequences to known routes, and thus identify transport modes and path alternatives taken by users. Using this methodology, they successfully reproduce the traffic demand in a mixed highway/rural setting in Baden-Württemberg, Germany.

A recent work by Uppoor et al. [92] evaluates pervasive mobile vehicular access to the cellular network in the TAPASCologne simulation scenario. The focus of this work is on the planning of the Radio Access Network with respect to vehicular connectivity. The authors studied cellular connectivity, dwell times and inter-arrival times, with respect to the Voronoi tesselation of the cellular network. The results show that these metrics exhibit considerable intra-day variability, which can be employed to improve the network infrastructure.

Our work in this study differs from these previous works in that we want to evaluate the feasibility of dwell-time-based travel time prediction using floating car data for validation purposes. We want to leverage the intra-day variability of dwell times as shown in [92]. Unlike previous studies [90, 93], our scope is country-wide, including both rural and urban roads and environments, and we focus on travel time prediction.

4.2 Estimating Travel Times with Synthetic Cell Dwell Times

As stated above, we create a synthetic CDT data set from FCD. In the following subsections, we show how the data was generated and how the graph-based model learns and predicts travel times. Finally, we present some validation results in different settings and provide additional findings and visualization techniques.

4.2.1 Data Set Description and Preprocessing

We base our analysis on a floating-car data set (FCD) with approximately 40 million datapoints that was gathered during a traffic monitoring campaign performed in Luxembourg in 2015. Floating-car data contains a vehicle's geographical position, bearing and velocity annotated with a timestamp and a vehicle identifier. We based our analysis on data from weekdays between Monday and Thursday, as these days exhibit similar traffic patterns [94].

The data set corresponds to a total of 27,124 trips in the Luxembourg area. Table 4.1 lists some statistics about the trips in our data set. Note that most of the trips are short,

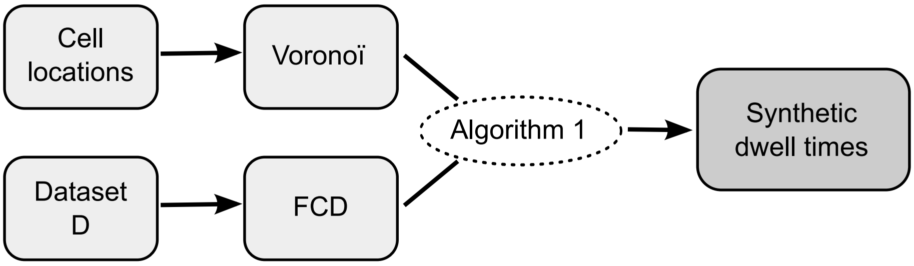


Figure 4.1: Overview of dwell time synthesis methodology.

with a median count of 8 visited cells and a mean trip duration of 2,122 seconds.

Overview

Starting from GPS data, we want to investigate whether cell dwell times could be an adequate means of estimating travel times.

In order to reproduce the cell dwell times of these trips, we propose to use the spatio-temporal preprocessing steps as described in the following subsections. Fig. 4.1 gives the outline of how we preprocess the Floating-Car Data (FCD) into synthetic cell dwell times.

The analysis was ran on a PostGIS-enabled PostGres database [95], into which we imported an OpenStreetMap export of Luxembourg map data using the osm2po tool[1]. PostGIS was used for the spatial coarsening of the data that we will explain in the following subsection.

[1] http://www.osm2po.de

Number of trips	27,124
Shortest trip duration	164 s
Mean trip duration	2,122 s
Longest trip duration	15,905 s

Table 4.1: FCD data set: Trip Statistics

Spatial Coarsening

In order to create a simple model of the mobile phone network, we computed the Voronoi tesselation with respect to the locations of the mobile phone base stations. For our test case of Luxembourg, there are 421 base stations of a single mobile phone service provider.

The Voronoi tesselation can be described as associating the coverage polygon to each base station. We coarsened the data spatially within PostGIS by associating the corresponding base station to every floating car datapoint, and subsequently dropping the GPS locations.

To summarize, the main purpose of this step is to transform the GPS trajectories into a sequence of visited Voronoi cells (i.e. the expected sequence of associated mobile phone base stations).

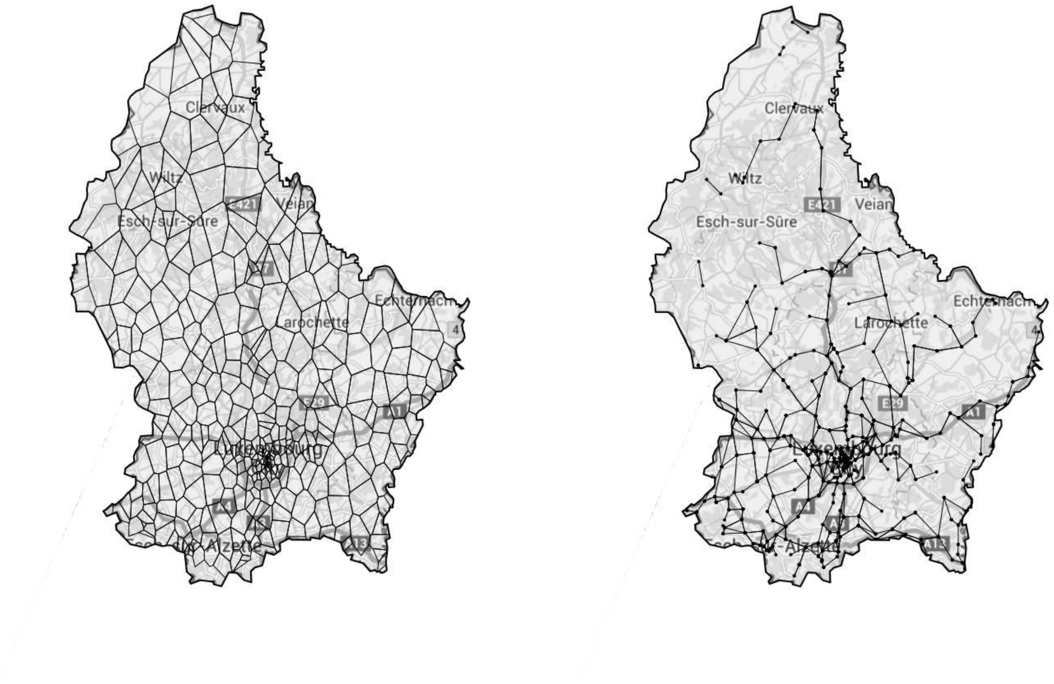
In Fig. 4.2a, we illustrate the tesselation of the country's territory with respect to the closest mobile phone cells overlaid on the Luxembourg road network [96]. Note that the cell density increases within urban areas, i.e. especially in Luxembourg City and Esch/Alzette (south). For travel time estimation purposes this means that a finer-grained resolution is possible within the areas where there are more points of interest. Fig. 4.2b shows a transition graph between the Voronoi cells, that we refer to in Section IV after we have introduce the dwell time model.

Fig. 4.3a shows the amount of handovers between 7:20am and 7:40am, i.e. during the morning rush. Note that most of our samples stem form highway traffic, especially from the ring around Luxembourg City. We are confident that due to the size of our data set, this is representative of the mobility in Luxembourg.

Fig. 4.3b shows the dwell times that resulted from the data preprocessing (algorithm 1). In order to filter trips with intermediate stops, we removed trips containing cells with dwell times over 500 seconds, i.e. 0.3% of the trips. We chose this threshold to remove outliers without removing trips exhibiting typical traffic jams.

Temporal Coarsening

We aggregate for each trip the entry and exit timestamps at each Voronoi cell (as defined above). Thus, we can compute the dwell time in this cell and associate it to an arrival time. The aim is to create a conditional distribution of dwell times with respect to different arrival times, and so as to keep the model simple, we group the arrival times



(a) Voronoi Diagram of the cell sites for the country of Luxembourg.

(b) Transition graph of Voronoi cells between 8:00am and 8:20am.

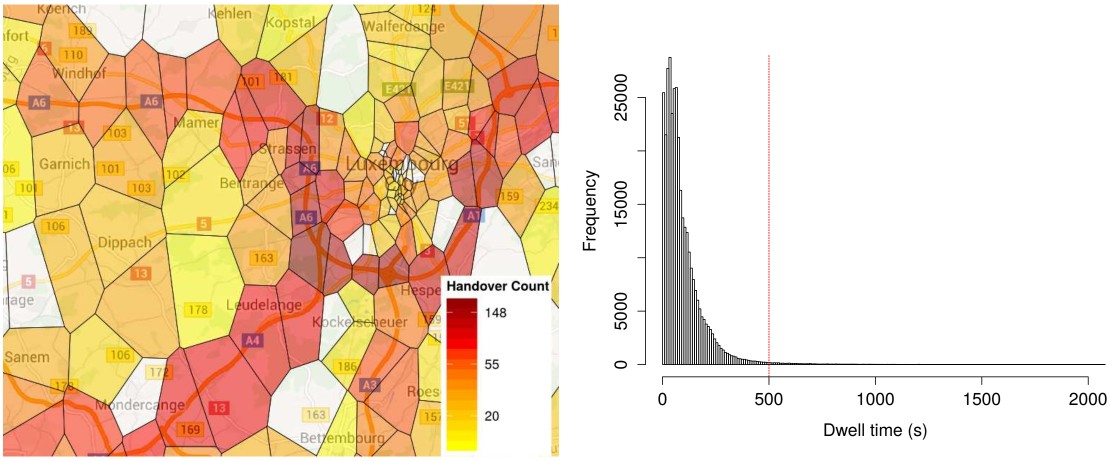
Figure 4.2: Voronoi tessellation and cell transition example for the country of Luxembourg.

temporally. We opted for 20 minute windows (temporal groups), as this value represented a good trade-off between precision and data availability. Thus, we coarsen the data to dwell times conditioned on approximate arrival times within 20 minute windows.

Note that we no longer look at the individual driver's travel times, but rather a distribution over the entire trip database. We further reduce the privacy footprint as there is no longer a need for information on the individual user.

Trip Truncation

In order to remove within-cell artifacts such as the search for parking spots, we truncated the trips at the first and last handovers respectively. Otherwise, the prediction error would be too sensitive to the exact start and end points within the cells. Furthermore, for



(a) Cell occupancy for the time slot 7:40am-8:00am (amount of handovers).

(b) Histogram of cell dwell times with the threshold value at 500 seconds.

Figure 4.3: Cell dwell time and cell occupancy summary statistics

the first and last cells visited, we have no source and destination cells to interpret in the training phase (i.e. no transitions).

Preprocessing Procedure

We summarize the steps described in this section in Algorithm 1, where we match the GPS beacons to the Voronoi tesselation, setting the source and destination cells, then drop all the samples that are not transitions and finally compute the dwell times.

With this preprocessed data set, we now have artificial dwell times and in the next section, we will proceed to perform travel time estimation using these dwell times.

4.2.2 Travel Time Estimation

In this section, we will introduce the dwell time representation as well as the travel time estimation procedure.

Synthetic Dwell Time Model

After grouping the data according to the coarsening measures above, we propose to capture the transition dynamics between cells using origin-destination (O/D) matrices. Note that we store both mean and standard deviation to compute our confidence intervals

Algorithm 1 Data Preprocessing

```
procedure PREPROCESS($FCDData\ D, Voronoi\ V$)
                                              ▷ Annotate Dwell Times in D

                                              ▷ set source cell
    for all sample $s \in D$ do
        $s_{source} \leftarrow s_{gps} \cap V$
                                              ▷ set dest. cell
    for all trip $T \subset D$ do
        for all sample $s \in T$ do
            $s_{dest} \leftarrow \text{successor}(s)_{source}$
                                              ▷ keep only handovers
    for sample $s \in D$ do
        if $s_{source} = s_{dest}$ then
            $D \leftarrow D \setminus s$

                     ▷ Set dwell times and truncate first and last cell
    for all trip $T \subset D$ do
        for all sample $s \in T$ do
            if $s_{dest} \neq \emptyset \wedge s_{source} \neq \emptyset$ then
                $s_{dwelltime} \leftarrow \text{successor}(s)_{time} - s_{time}$
            else
                $D \leftarrow D \setminus s$
     return $D$
```

under the assumption of a normally distributed error. Since there is no covariance model between cell transitions, we consider them as independent in the calculation of a route's standard deviation. The evaluation section will show that these assumptions hold (i.e. that the confidence intervals are of use) for at least one of the three route selection methods.

For our analysis, we computed 72 sparse O/D-matrices, i.e. one for every 20 minutes of a typical work day (we considered Monday until Thursday [94]). These matrices represent the distribution of dwell times observed between origin and destination cells given an arrival within the 20 minute window. The sparsity is due to the fact that only neighboring cells (between which a handover can take place) have transitions.

We will now introduce the notation to describe the cell dwell time graphs and the prediction equations. Let i be the number of base stations in the prediction scope and $t \in [0, 71]$ be the 20 minute time slot index. Let $\mu_t^{i \times i}$ be the sparse matrix of mean dwell times between origin/destination base station pairs, and $\sigma_t^{i \times i}$ contain the corresponding matrix of dwell time standard deviation. Assuming a sequence of cell transitions *path* (of length $n_{transitions}$) as output of a routing method between the origin and destination of our prediction request, we have the following prediction rules:

$$\text{Travel Time: } TT_{path} = \sum_{(a,b) \in path} \mu_t^{a,b} \tag{4.1}$$

$$\text{Standard Deviation: } SD_{path} = \sqrt{\frac{\sum_{(a,b) \in path} (\sigma_t^{a,b})^2}{n_{transitions}}} \tag{4.2}$$

In the remainder of this study, we will be using these equations to compute the travel time estimations and confidence intervals.

Fig. 4.2b shows an example transition graph corresponding to μ_{25} (time slot of 8:20am-8:40am) connecting the different mobile phone cells overlaid on top of the Luxembourg road network. As expected most of the edges follow the major roads and highways.

In order to model the dwell times, we opted for a univariate normal distribution, thus we compute (μ, σ) for each O/D-pair, i.e. each conditional dwell time. We chose this solution as we are not directly interested in the distributions themselves, but rather in the mean transition time. Otherwise, it would have been necessary to fit different probability density functions to our dwell time data, e.g. the hyperexponential distribution [97]. While in many cases this would have been of high importance, choosing the ideal distribution would not have provided much added value for this study. We only need the expected value and an approximate measure of standard deviation to see whether travel time prediction using dwell times is feasible. In a later implementation, it will of course be useful to fit more appropriate distributions when performing predictions on real data.

It is important to stress that in this study, due to the nature of our data, we can only evaluate travel time predictions based on a biased sample of dwell times (only vehicles). However, we are confident that a sensible cut-off value over the full dwell time distributions (including e.g. pedestrians and users that are not in transit) will yield comparable results, as it was the case in [90], where the authors succeeded in separating dwell times corresponding to road traffic. Furthermore, most new cars are equipped with SIM-cards, e.g. because of the eCall system. By looking only at data from this category of mobile equipment, we could obtain purely vehicular mobile traffic data. For this approach, the question of feasibility is more dependent on cost than technical barriers.

Prediction Procedure

The estimation of travel times using this model follows a simple procedure as listed in algorithm 2. First, we evaluate which cells are visited on the trip from source to destination, i.e. the routing step. Then, we use the transition matrix μ_t corresponding to the departure time slot t and evaluate equations 1 and 2 to obtain estimates of the path's travel time and its standard deviation.

Algorithm 2 Travel Time Estimation

```
procedure PREDICT($t_{leave}, cell_{origin}, cell_{dest}$)
    $path \leftarrow route(cell_{origin}, cell_{dest})$
    $t \leftarrow \frac{t_{leave}}{1200}$
    $TT \leftarrow 0$
    $varsum \leftarrow 0$
    for all ODpair $(a, b) \in path$ do
        $TT \leftarrow TT + \mu_t^{a,b}$                          ▷ eq. 1
        $varsum \leftarrow varsum + (\sigma_t^{a,b})^2$
    $SD \leftarrow \sqrt{\frac{varsum}{len(path)}}$              ▷ eq. 2
return $TT, SD$
```

In the following section, we will present the results of the evaluation of the model described above.

4.2.3 Evaluation

For evaluation purposes, we split our data set into training (70%) and test (30%) sets, using random sampling without replacement. We compared the prediction results with the ground truth floating-car trip data. Fig. 4.4 shows the distribution of trip lengths (expressed in minimum amount of cells visited) in the test data set after the preprocessing

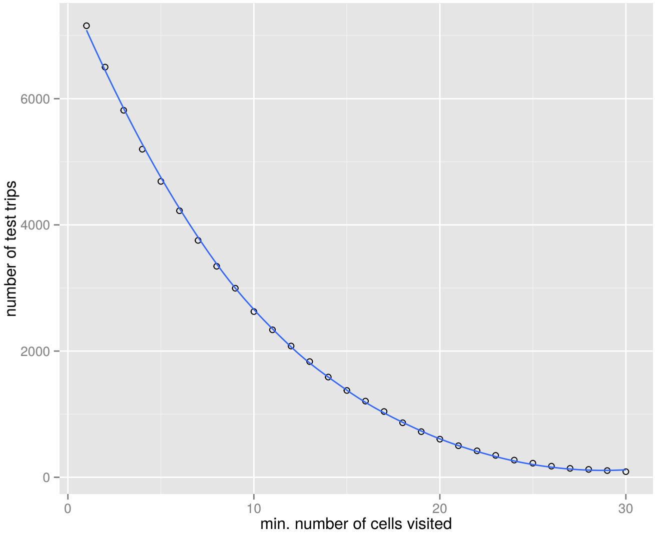


Figure 4.4: Number of test trips relative to the trip length in cells visited and locally weighted regression line (blue).

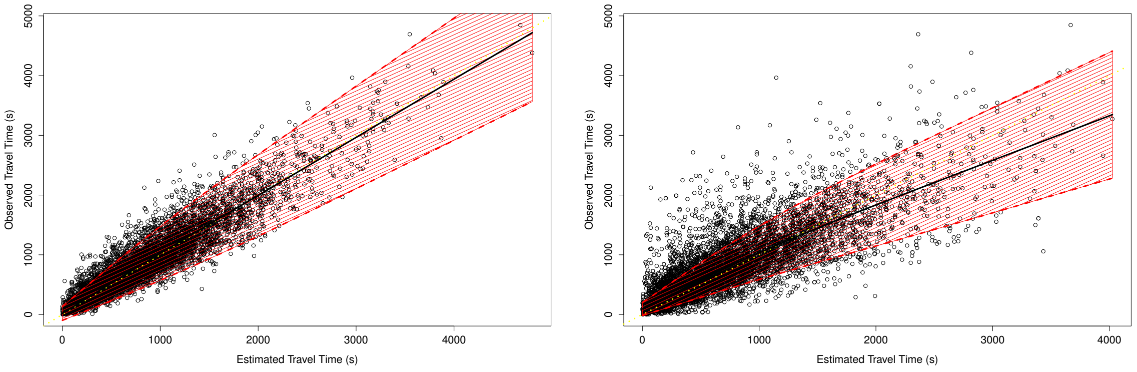
step that we describe in the following subsections. Note that most trips are short with a median cell sequence length of 8 and a mean trip duration of 2,122 seconds (as mentioned in Section III).

It is important to note that not all of the trips in our data set corresponded to simple one-way trips, but that some of them included detours and short stops. Thus, some of the estimation error below is due to these *indirect* trips, i.e. in the situations where the cell sequence was masked (unknown trajectories). For this reason, we refer to the median error in Fig. 4.6, as some of the indirect trips impact the error percentage by a large margin because the routing algorithms obviously find direct paths (cf. subsection 4.2.3 for more details).

Results: Known cell sequence

We have compared the estimated dwell times against the original trip traces, in order to evaluate the error.

The error can be approximated by a normal distribution, as can be seen in Fig. 4.7, but we also observe a slight tendency of underestimating travel times, which can be due to traffic jams that go beyond the typical slowdowns reflected in the transition matrices.



(a) Known cell sequence

(b) Unknown cell sequence

Figure 4.5: Estimated and observed travel times 95% confidence intervals.

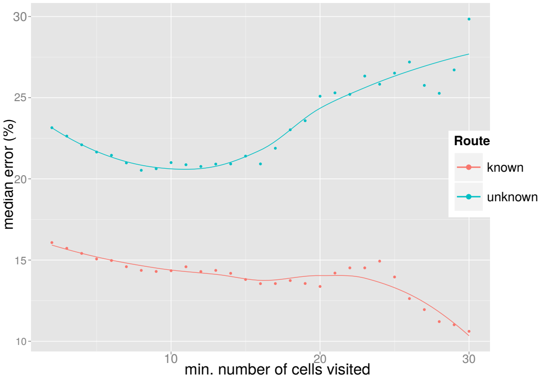


Figure 4.6: Median error percentage relative to trip length in number of cells visited and locally weighted regression lines.

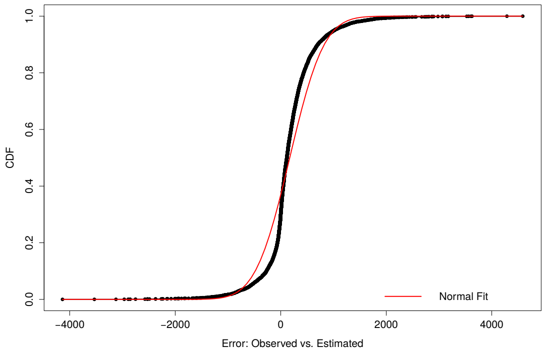


Figure 4.7: Empirical CDF of the error distribution and Gaussian fit

The per-cell variability of dwell times in a trip makes for a larger relative error in short trips as shown in figure 4.6. As indicated in the figure, the relative error drops with growing trip length, as the impact of small perturbations (e.g. traffic lights) is cancelled out and the error drops to about 10.5%.

Generally speaking, we can see that there is a good match between the travel time predictions with their respective standard deviations and the observed data, even for the longest test trips of around 5,000 seconds.

Results: Unknown cell sequence

We also evaluated the model's performance when removing the knowledge of the intermediate cells, instead using the observed paths from the training data. Thus, the only input data remaining are the origin and destination cells as well as departure time, i.e. the typical minimum inputs of a routing request with travel time estimation. We evaluate all the path alternatives present in the training data set, and choose the fastest one for our prediction. This reflects the typical user behavior of avoiding crowded paths (e.g. during rush hour).

Fastest observed path

At first, we tried using only the fastest road paths, which we precomputed using pgRouting. We observed that many estimated trip times differed strongly from the observed trip times due to the difference in path taken. Users often take different paths to avoid traffic at different times of day, and these do not correspond to the shortest path.

Therefore, the next step was to remove the requirement of knowing intermediate cells. Thus, we wanted to use only origin and destination cells to determine the path travelled and predict travel times. In order to achieve this, we learned the paths connecting O/D-cell-pairs from the training data.

Fig. 4.5b shows the results for this type of prediction. While we can see that in the majority of cases, the test cases fell within the 95% confidence interval, it is also clearly observable that there is a significant proportion of trips, especially short ones, that are over- or underestimated. This is either due to the test path being indirect (as defined above) or the fact that (especially for the long paths) a different path was found. Thus, if the duration was underestimated strongly, this indicates the availability of a faster path alternative to the one the user actually took.

In Fig. 4.6, we can see that for this case (unknown route), the median error is generally higher than in the known route case. As mentioned above, this is partially due to indirect trips in our test data set. The other effect that we can observe is that with increasing trip length (expressed here in terms of number of cells along the path), there is an increase in the error. We attribute this increase to the likelihood of finding a different path than the one taken by the user, which increases with the length of the paths, as the number of alternative paths grows.

Dijkstra

Finally, in order to see if a graph routing algorithm can provide valid results, we used the Dijkstra algorithm to find for each O/D-pair the fastest path using the current transition time matrix.

We compared the cell sequences found by running the Dijkstra algorithm on the transition matrices with the actual user cell sequences. Fig. 4.8 shows the distance in number of cells visited between the observed test data paths taken and those suggested by the Dijkstra algorithm. The positive mean (full stroke red line) shows that observed paths were longer on average in terms of cells visited. The large standard deviation of the differences indicates that the paths found differ in the vast majority of test cases.

Indeed, our results indicated an inacceptably large mean error for this approach. The error was mostly one of underestimation, i.e. the case where the Dijkstra algorithm finds a seemingly faster path through the road network than was used on the actual trip. Contrarily, in cases where the travel times were overestimated, this underestimation was mostly due to the sparsity of data for some times of day (e.g. missing links in the adjacency matrix of the graph at night).

Our conclusion is that one needs to be careful when relying on the Voronoi representation for routing. An existing transition in the graph does indeed show that there is a road link, but it does not guarantee that this link can be reached from the road segment taken on the previous transition. We will discuss this point and a potential solution in the following subsections.

Limitations

The results in this study are subject to two kinds of limitations: On one hand, a Voronoi tesselation is no replacement for a real propagation model, but it is commonly accepted in mobile traffic analysis as a first approximation. On the other hand, we have only been able to model dwell times of cars, whereas with full-population data, we would have to work with different (phase-type) distributions, and decide on what quantiles to consider for travel time estimation purposes. However, as mentioned previously, there are special types of car fleets with mobile equipment that could be more easily identified and used as a ground truth for establishing these rules.

We have observed different behaviors and prediction errors with the 3 route selection methods. Unsurprisingly, following the exact cell sequence (the known route) delivered

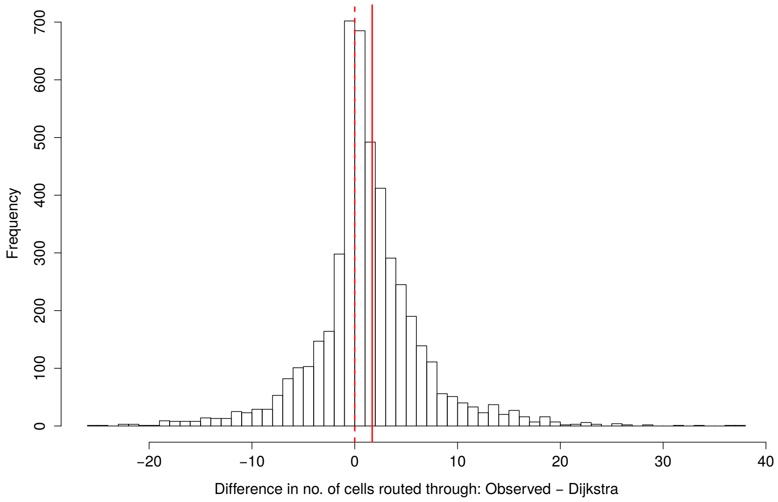


Figure 4.8: Distribution of the path length difference using Dijkstra.

the best results, as the prediction routine will always follow the user path. Using the data-driven best observed route method yielded usable results, that were partly worsened by indirect paths taken by users (i.e. paths with intermediate stops of short duration). Removing paths with dwell times over 500 seconds (which made up 0.3% of dwell times) proved insufficient to account for some of the short stops, as further inspection showed some trips still contained stops. However, simply reducing this threshold would also mean cutting out direct trips with traffic jams. Therefore, over a larger data set, it would be sensible to cluster O/D-trips into common paths and to remove the anomalous and much longer trips, as it was proposed by Giannotti et al. [98]. On the upside, it was also possible that the best learned route would have been faster than that taken by the user on their actual trip. However, this approach in practice would require more data than just the dwell time distributions, as it would require tracking some users to find out the real cell sequence patterns along the paths.

The Dijkstra approach also proved to be problematic. On one hand, if there is insufficient data, we cannot guarantee finding a path in the first place, as the transition matrix might be too sparse to find one. However, there is also the problem that the algorithm can potentially find transition sequences that are not feasible on the road network, e.g. if one transition would require changing from a highway to a secondary road and there is no exit within the current cell.

Summary

We found that the results along a known route delivered the best results. The disadvantages of using data-driven or graph-based routing were too important to neglect, and the results were unsatisfactory. Therefore, in a practical travel planner setting, we recommend using a routing algorithm such as Dijkstra or A* on the road network to provide multiple path alternatives, and use the presented model to evaluate the travel times along these paths' cell sequences as a cost function. This way, one only needs dwell time data and the road topology and can provide optimized (potentially real-time) path recommendations.

As a potential addition, the Dijkstra approach with road network verification can be added: The road network intersection with the Voronoi cells in the Dijkstra path is computed and a routing algorithm is ran on this set of road segments. This allows to see if there is a drivable, faster alternative, that is not necessarily short in terms of road distance and potentially was not considered fast by a road-based routing algorithm (which typically would be using free-flow travel times as a cost function). Otherwise, it is possible to proceed as described above, i.e. with normal road network routing and travel time estimation using the dwell time model. Thus we suggest, for real applications, to try out a combination of these strategies to find the fastest possible road path, joining the knowledge of road topology and the transition graph as a cost function.

Furthermore, scalability can be achieved through the regrouping of zones. The sparsity of the transition matrices should, however, only make this necessary for very large prediction scopes.

Generally speaking, our results indicate that signalling data could indeed be used to make travel time predictions at a large scale, as long as we manage to identify mobile users (e.g. drivers, passengers, etc.).

4.2.4 Visualization

In order to represent the model in a comprehensible way, we will introduce in this section a novel flow-density graph. We want to use both the model's spatial regrouping of trips as well as the information on handovers that we estimated.

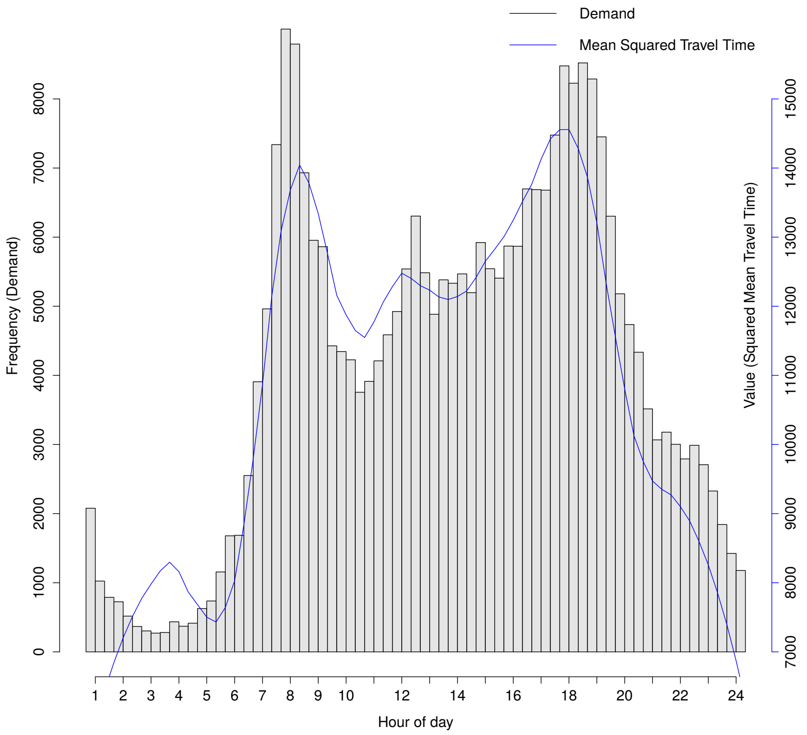


Figure 4.9: Demand (number of handovers) vs. travel time.

Relation between Handover Count and Dwell Times

In Fig. 4.9 we can observe a proportionality between the amount of handovers and the squared average dwell time (i.e. the squared mean of the entire O/D-matrix):

$$n_{handovers} \propto \overline{t_{dwell}}^2 \tag{4.3}$$

It is subject of future work to determine, at which scale this proportionality holds, and whether it concerns rural and urban roads equally. Leveraging this relationship, we introduce a novel manner of displaying traffic using our graphical representation. We visualize simultaneously the estimated handover (and thus vehicle) density and the estimated slowdown of cell transitions (traffic slowdown). This type of plot aims to indicate where traffic anomalies arise and in which regions people are the most active.

We define the flow slowdown as the coefficient of the transition time with respect to the daily mean.

Example Plots for Luxembourg

As a means of reference, we show the population density in Luxembourg in figure 4.10. The central high density zone corresponds to Luxembourg City, while the south-eastern high density zone corresponds to Esch/Alzette. Note that the population density correlates to the increased handover we estimated in the following flow-density figures.

The first flow-density figure 4.11 visualizes the flow slowdown and estimated handover density in the morning rush peak hour, i.e. in the 7:20am-7:40am timeslot (see also Fig. 4.9). It can be seen that significant slowdowns are observable on the main axes pointing towards the center of the map, i.e. Luxembourg City. They are due to (cross-border) commuters moving towards the main working areas within Luxembourg City. In the very center of the map, we can also observe the swirl motion of flows moving around the city center in peripheral business districts.

Fig. 4.12, by contrast, shows the off-peak flow slowdowns in the 12:40pm-1:00pm time slot. Note that slowdowns are both less severe and less frequent, while handovers are estimated to happen much more evenly, i.e. mobility is less centered around the business areas.

Fig. 4.13 shows the evening rush hour for the 5:40pm-6:00pm time slot. Here, we can observe the flows leaving the capital city over all peripheral highways. This highlights the star-topology of flow directions in the country and the challenges the infrastructure faces on a typical workday with a slowdown factors of up to 5.

When supplying these diagrams with real-time expected values of conditional dwell times, it is possible to identify travel time slowdowns visually.

4.2.5 Discussion

In this study, we have evaluated whether mobile phone dwell time data is suitable to model car travel times. We have shown that a simple, time-discrete graph can predict car travel times with sufficient precision for many applications. We have also shown that spatial and temporal aggregation of data allow privacy-neutral travel time estimation, that could be provided as a service by mobile phone carriers. Furthermore, our results indicate that using alternative routing schemes could provide users with faster travel routes when combining knowledge of the road topology with cellular dwell times. Finally, we have introduced a novel representation to visualise traffic flow slowdowns and handover density allowing to observe a potential correlation between the two metrics.

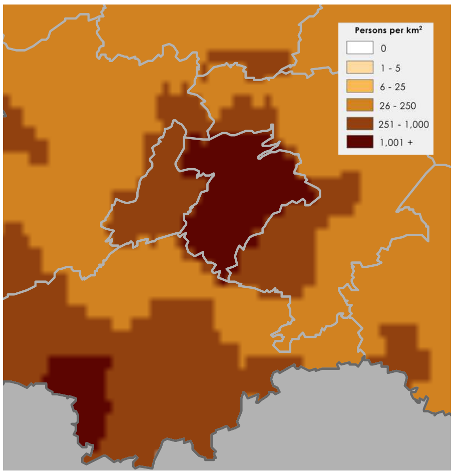


Figure 4.10: Population density: South of Luxembourg [99].

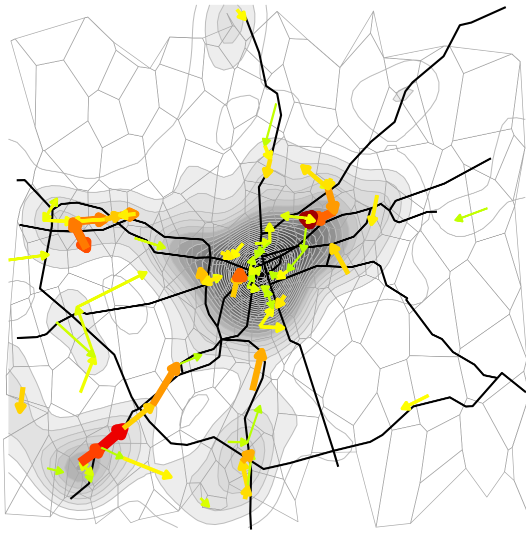


Figure 4.11: Flow-density graph: 7:20am-7:40am (morning peak).

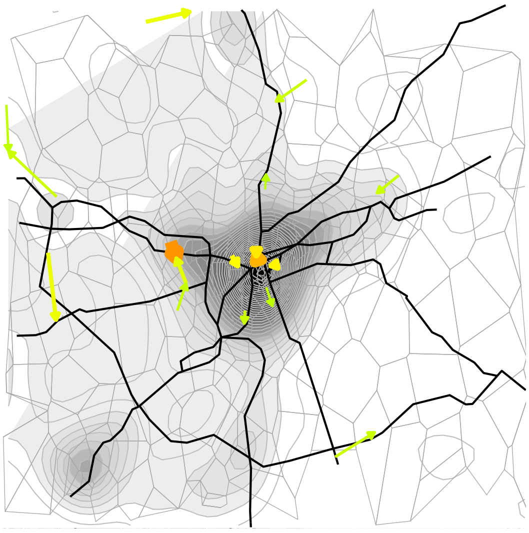


Figure 4.12: Flow-density graph: 12:40pm-1:00pm (off-peak).

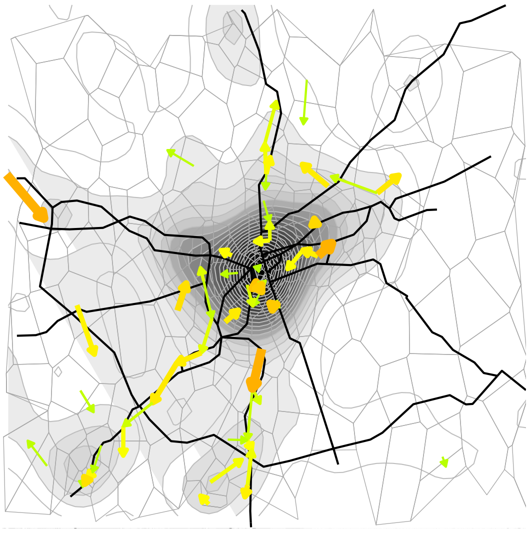


Figure 4.13: Flow-density graph: 5:40pm-6:00pm (evening peak).

Chapter 5

Supporting Demand Modeling with Mobile Network Data

Intuitively, mobile network data has great potential for improving demand estimation, as users appear in the data based on where they are. In this chapter, we want to show ways of supporting demand estimation models using mobile network data. We suggest methods of imputing user trajectories in incomplete data, in order create the temporal component for synthetic population generation. This is a complementary work to that of Di Donna et al. [100], who presented a CDR-based Markov Chain of user movements, that can provide the corresponding spatial component of the synthetic population model. Applied to demand estimation, the estimated CDT distributions can be used for likelihood computation of activity durations in activity-based demand estimation methods.

Next, we show how aggregated handover counts can be used in an activity-based demand modeling framework. Thus, we want to show that mobile network data can serve for conditioning purposes to improve the reliability and convergence of existing demand estimation schemes.

5.1 State of the Art

In activity-based demand modelling, estimating user locations and the duration of their activities there are a common approach to reproducting their intermediate mobility and mode choices. With respect to mobile traffic data, activity locations can be modelled through the distribution of visited cell sites. In this context, Halepovic et al. [101] found that for the majority of mobile phone users, mobility is sparse and that a small, but non-negligible percentage of the population exhibits a high mobility behaviour, i.e. the distribution of user mobility is heavy-tailed. Subsequent studies have confirmed this trend [102], [103]. The work of Kung et al. [104] studies the commuting behaviour between home and work locations using, among others, CDR data sets. In order to identify the home and work location they first constructed and *Individuals' Travel Portfolio*, which consists of a list of frequently visited cells. This and other parameters are then used to identify the home and work cells. These results can be used for the generation of demand models of human mobility, which are typically expressed using Origin-Destination Matrices.

Tizzoni et al. [105] show how census data and mobile network data can create similar demand and thus mobility patterns, and used the results in an epidemiological study on the countries of France, Spain and Portugal. They emphasize that mobile traffic data is important for studying disease spread in countries where census data is unavailable. Zhang et al. [106], on the other hand, show how mobile traffic data can be enriched in open-data environments. They augment CDR data with floating-car data from Taxi and Bus fleets and add Smart Card Data from Buses and Subway. They use this rich data stream to perform online inference on the amount of flows between origins and destinations, showing significant estimation improvements over CDR data alone.

Another interesting application of CDR-based demand models is traffic optimization via iterative simulations. Zilske at al. [9] show that mobile phone transactions without any layer of interpretation provide plausible traffic patterns. It has however been pointed out that further verifications would be needed to validate their assumption. For simplicity they used the great circle distance to estimate the commuting distance between the home and work location. Along with demand estimation, the analysis of cell dwell times, i.e. the time a user spends at one location, provides important insights on human mobility. Over the past decades different probability distributions have been used to model cell dwell times [107]. Fang [108] found that *Phase-Type* (PH) distributions provide a very good description of the dwell times. Among PH distributions, Coxian and Hyper-Erlang provide the best fit due to their universality property. Similarly, in [109], the authors found that there is a relationship between channel holding time, i.e the time a user remains in the same cell during a call, and cell dwell time by taking the assumption that cell dwell times are Coxian or Hyper-Exponentially distributed. They also presented *Extreme Value Distributions* (EV) for modeling cell dwell times, such as the Generalized EV Distribution.

Hidaka et al. [110] perform an analysis on the cell dwell times observed in a taxi fleet. They found that in their data set, the cell dwell time distributions can be approximated by a log-normal distribution. All the previously mentioned distributions are heavy-tailed and can express both transition and activity cell dwell times of network users. There is however no single distribution that works best in all use cases, as the dwell time distributions depend on a country's infrastructure.

In [111], the authors propose for example to infer transportation modes based on CDRs. They determine, for a given origin and destination, the percentage of travelers using each transportation mode based on their travel times (e.g. walking, public transit, driving cars). The authors of [112] use CDRs generated from mobile phones in Tallinn (Estonia) to estimate the composition of traffic flows. In [113], the authors focus on characterizing the population living in dense urban areas from CDRs using a Voronoi tessellation to define the coverage area of cell towers.

Finally, the work presented by Apolloni et al. [114] at the D4D Ivory Coast Challenge provide some interesting insights on the limitations regarding the generation of synthetic populations using only CDR data. They make the assumption that the inter-site time is proportional to the lengths of the intersections of the straight line between the two cell sites. Further, they used simple assumptions regarding the composition of households and mobility patterns to build a demand model. Due to the lack of alternative data sources their results could not be validated.

Toole et al. propose a full OD-estimation framework based on CDR, geographical and census data [115]. By contrast, we focus on generating cell dwell time distributions, thus obtaining a generative conditional model of activity durations in different locations.

In the methodology used in this study we introduce two new *attractivity* metrics to characterize the mobility of both an individual user and the entire population, combining some of the ideas from [104] and [105]. The motivation is to estimate dwell time distributions during periods where no, or limited data is available (e.g. between two consecutive network activities), enabling the construction of a synthetic population. In contrast to [114], we add knowledge of the road topology and the attractivity metrics to produce realistic dwell times from CDR data. Our approach defines a first step towards a realistic synthetic population that can be used as an input to perform traffic simulations and optimizations, while relying on publicly available roadmap data to better estimate distance and travel time between distant cell sites.

5.2 Synthetic Cell Dwell Times from Call Detail Records

Knowing the geographical location of cell towers allows the reconstruction of users' spatio-temporal activities on a countrywide scale, providing important insights into human mobility. This is especially relevant in developing countries, as alternative data sources are often very limited. Further, the mobile phone penetration in these countries is usually very high, which implies that these data sets are representative of their populations. As an example, the mobile phone penetration rate in Senegal recently exceeded 100% [116].

One limitation of CDR data sets is that they only provide information about the location of a user when a transaction is being made, e.g. during call or messaging activity. It is therefore interesting to estimate the spatio-temporal paths travelled by users between their consecutive appearances in the CDR data, the *path imputation problem*. The state of the art work was carried out by Apolloni et al. [114] in the context of the D4D Ivory Coast Challenge, who provided some interesting insights into the generation of synthetic populations using only CDR data. The limitation of their approach is that they consider unrealistic paths that are independent from the underlying road topology. This method also does not take into account places of interest and user behavior.

Two other major limitations in the use of CDR data sets are privacy concerns due to the traceability of (pseudonymized) customers and the typical size of the data sets in the order of multiple gigabytes of data per day [103, 117], even in aggregated form over several months [4].

In this article, we address the limitations mentioned above, allowing the network dynamics to be described in a more compact, privacy-neutral manner, and enabling mobile networks operators to share them. More specifically, we present a novel methodology for path imputation and cell dwell time estimation that improves upon the aforementioned approach by Apolloni et al. [114]. We model the trajectories as the fastest road paths between consecutive recorded activities. This is achieved through a combination of OpenStreetMap data used for the routing, and a Voronoi tesselation of the base stations to determine the visited cell sequences.

Further, it is necessary to weigh the proportions of time spent in the different cells along the path between consecutive activities. In our proposal, we introduce two *attractivity* metrics to weigh the importance of locations both for individual users and the entire user base. As in [104], we separate day- and nighttime activity to identify users' work and home locations, respectively. By combining the aforementioned metrics in a weight

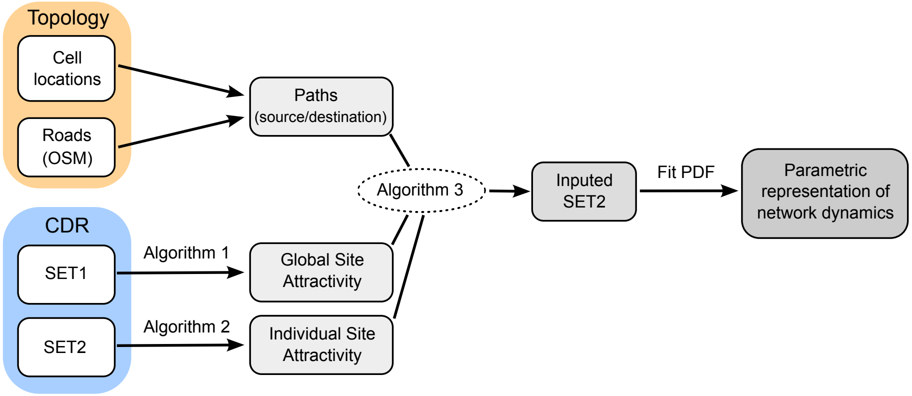


Figure 5.1: Synthetic CDT from CDR: Methodology overview

function, it is possible to estimate dwell times between CDR activities, creating full user trajectories. From these full trajectories, we can then group the dwell times by base station conditioned on arrival time and destination base station. To each group, we then fit the parameters of conditional probability distributions (which can then be shared and are privacy-neutral). Using this model, we can reconstruct user trajectories from CDRs by imputing intermediate locations and times of arrival. Our analysis is based on the *Data For Development* (D4D) Challenge data set from 2014, which consists of CDR data provided by Orange for the country of Senegal. This model is to be seen as the temporal component for synthetic population generation, while the spatial component could be generated through a Markov Chain model, as presented by Di Donna et al. in [100].

The remainder of this article is structured as follows. In Section II, we present our model and introduce the necessary notation. In Section III, we describe the D4D Senegal data set [4] and perform a numerical case study using our model on its. Finally, in Section IV, we provide a discussion of our results, comparing them to other data sets and related work, and conclude in Section V with directions for future work.

5.2.1 Methodology and Notation

This section introduces our methodology for path imputation and dwell time estimation in CDRs, along with the necessary notation. Note that the presented steps are intended to be performed by mobile network operators (MNOs) to compute a compact, privacy-neutral model representation of the mobility in their network. Figure 5.1 shows an overview of the methodology. Note that we use the road and network topology as inputs, as well as the CDR data sets that we describe in the following subsection. We perform path imputation from these inputs in Algorithm 3 to construct the imputed SET2, from which we extract the parametric representation of the network dynamics by fitting probability density functions (PDFs) to the imputed dwell times. In the remainder of the section, we introduce our assumptions on which we build our model. We then introduce attractivity CDR metadata, which allows the imputation of user positions in CDR data sets, along with the necessary Algorithms 1-3.

Necessary Input Data

In order to model cell dwell time distributions for a given mobile network, we propose to use two data sets that can be made available by the network operator, typically through the billing system in the form of CDRs, as provided in the D4D Challenge data sets [4]. We go into further detail regarding these data sets in the numerical case study in Section III.

- **SET1: Aggregate data set** of hourly traffic between mobile antenna pairs – call and SMS activity of the entire customer base
- **SET2: data set of individuals**' consecutive call and SMS locations for a small subsample of the mobile customer base

Assumptions

We will now introduce three assumptions necessary to build our dwell time model. They underlie the rules that allow the inter-activity times to be systematically split.

Assumption 1. We assume that the amount of call activity in a site reflects the amount of people and mobility present in that area. That means that the probability ($P_t(s)$) of a user being located in a site (s) within a time window (t) is proportional to the aggregate amount of call activity in that site during t:

$$P_t(s) \propto A_G(s,t) \ , \tag{5.1}$$

where the aggregate amount of activity is denoted as global (objective) site attractivity (A_G) and is computed using the metadata generated using Algorithm 3,as described in detail in subsection 5.2.1.

Assumption 2. We also introduce a measure of individual (subjective) attractivity, i.e. how often a user has performed a call or messaging activity from a site. This serves to increase the estimated dwell time in frequently-visited locations, such as home and work. Such a measure is required because pseudonymized CDR traces are generally not available for the entire population, but only a subsample for which we want to identify points of interest (POI). We distinguish between daytime and nocturnal POIs and identify them using the metadata obtained from Algorithm 4. The following expression translates the idea behind this assumption, i.e. that a user is more likely to remain at a location where he/she more frequently performs network activity:

$$P_t(s) \propto A_I(s,t) \tag{5.2}$$

Note that these assumptions influence the dwell time distributions. A_I in particular is used to create a certain bias for users' favourite locations, e.g. home and work, and to make sure that we allocate longer dwell times to these locations. Some transportation models allow for dynamic route choice. This allows the movement of some individuals without using the fastest path (e.g. a less occupied path to the same destination). However, it is not in the scope of this study to use randomized route and mode choices. Using these basic assumptions, we will describe in the next section the exact computation of the attractivity metrics A_G and A_I.

Assumption 3. We also assume that users travel on the shortest road network path between two consecutive call locations. During our research on the Senegalese transportation infrastructure, we found that the prevalent means of transportation is via the road network using private cars, buses or taxis[1]. While this is a significant assumption, we believe that it is the obvious and most reasonable choice in the absence of ground truth regarding user movement between cells. Its effect is mitigated on an aggregate level by the two previous assumptions, since cells with high activity are allocated more of the inter-activity time of users, thus reducing the importance of the path taken between them.

Assumptions 1 and 2 are needed to perform the temporal split between activities, and Assumption 3 is needed for the spatial reconstruction of the inter-activity paths that we estimate users took.

SET1 and SET2 Data Extraction

In order to create realistic movement patterns, it is necessary estimate the intermediate movements of users between their call and messaging activities. Having computed

[1] http://worldbank.org/transport/transportresults/regions/africa/senegal-output-eng.pdf

routes between consecutive activities, we now propose to combine our knowledge about the general population from the aggregated data set (SET1) with the knowledge about individual users (SET2) to estimate the dwell times along the routes. We now introduce the necessary pre-processing of the data sets, which creates the metadata for computing the attractivities A_G and A_I of the different sites.

For SET1, the precomputations involve calculating the total call count and duration at each site and hour of day (cf. Algorithm 3).

Algorithm 3 SET1 Global Attractivity Computation

```
group data set by $site_{source}$, $hour$
for all [$site_{source}$, $hour$] do
    aggregate $callcount$, $callduration$
```

For SET2, successive user locations and timestamps are put into a sequence of tuples of $[location, timestamp]$ (a zip-list) of each user in Algorithm 4. This set of metadata allows us to quickly count which locations were the most frequented by a user, which we need for the computation of the per-individual site attractivity that is introduced in subsection 5.2.1.

Algorithm 4 SET2 User Trajectory Extraction

```
for all $userid$ do
    group by $userid$
    aggregate $ziplist(siteid, timestamp)$
```

Attractivity Computation

The attractivity measures serve to estimate the amount of time a user has spent within a certain site. We propose two different measures, i.e. the global attractivity (spanning all users of the network) and the individual attractivity (concerning a single user). The purpose of these metrics is to enable the combination of both the behaviour of the whole population, and the individual behaviour of the the user. According to the assumptions from the previous section, we propose the following attractivity measures:

Global Site Attractivity

The global attractivity of a site in a given hour of day (see Equation 5.3) represents the overall call activity that exists in this site, i.e. the sum of the number of calls (n_{calls})

to all destination antennas (d) from a given site (s) and in a given hourly time slot (t). We base this computation on the data obtained from the precomputation Algorithm 3.

$$A_G(s,t) = \sum_{d \in Antennas} n_{calls}(s \to d, t) \tag{5.3}$$

The A_G metric gives us a measure of the importance of a site at a given time with respect to the entire network.

Individual Site Attractivity

The A_I metric gives us a measure of the importance of a site at a given time of day with respect to a single user.

We distinguish night and day attractivity according into two different time categories. We consider the daily attractivity of a user (u) at a site (s), the number of activities (n_{act}) between 7:00 and 21:59. The nightly attractivity corresponds to the amount of activities between 22:00 and 6:59. We retrieve n_{act} from the metadata generated earlier using Algorithm 4.

$$A_I(u,s)_{Day} = \sum_{t \in [7{:}00 - 21{:}59]} max(1, n_{act}(u,s,t)) \tag{5.4}$$

$$A_I(u,s)_{Night} = \sum_{t \in [22{:}00 - 6{:}59]} max(1, n_{act}(u,s,t)) \tag{5.5}$$

Depending on the starting time of the two successive activities, between which we want to impute intermediate locations, we choose the corresponding attractivity measure. Note that we normalize all the attractivity measures, in order to use them as weights for estimating the times spent in intermediate sites.

Dwell Time Model and SET2 Imputation

In the numerical case study (section 5.2.2), we explore different functions F combining both attractivity metrics (A_G and A_I) into weights that split up the inter-call times into dwell times. They are used in Algorithm 5 to build an imputed SET2: We iterate over all user's trajectories and each pair of successive user activity cells. The travel route between them is queried from the spatial database, i.e. a sequence of cells on the road path between them ($sites_{intermediate}$). Then, both attractivity metrics are evaluated for all of these sites

Algorithm 5 SET2 Trajectory Completion

```
for all [$user$, $trajectory$] in data set do
    $traj_{full} \leftarrow [\,]$
    for all [[$site_1$, $t_{site_1}$], [$site_2$, $t_{site_2}$]] in $trajectory$ do
        $\Delta T \leftarrow t_{site_2} - t_{site_1}$
        $sites_{intermediate} \leftarrow$ CellsBetween($site1, site2$)
        $Dwell \leftarrow [\,]$
        for all $site$ in $sites_{intermediate}$ do
            $A_G \leftarrow$ GetAttrGlobal($site$, $hour$)
            $A_I \leftarrow$ GetAttrIndiv($user$, $site$, $hour$)
        $Weights \leftarrow F(A_G, A_I)$
        $Weights \leftarrow Weights / \sum_{Weights}$
        for all $site$ in $sites_{intermediate}$ do
            $Dwell.append(Weights_{site} \times \Delta T)$
        $traj_{full}.append(ziplist(sites_{intermediate}, Dwell))$
    $Output$ [$user$, $traj_{full}$]
```

and the time difference ΔT is split according to the attractivity weights computed using a weight function $F(A_I, A_G)$, the formulation of which is subject to the model selection that is performed in section 5.2.2.

Finally, we obtain the dwell times for each individual site and can compute the arrival times in sequence by summing up the dwell times. This yields the fully imputed user trajectories.

Dwell Time Model

The temporal split of inter-activity time ΔT spanning over an inter-antenna route ($path$) can be decomposed into dwell times d for each segment s of $path$:

$$d_{s \in path} := \sum_{t=t_{start}}^{t_{end}} \frac{F(A_G(s,t), A_I(s,t))}{\sum_{p \in path} F(A_G(p,t), A_I(p,t))} \Delta T \tag{5.6}$$

This equation links the SET1 and SET2 information (A_G and A_I) into the dwell time split through the F function, which we explore in the model selection section, 5.2.2.

To summarize, we have established a linear combination based on both attractivity metrics that decomposes the inter-activity time into separate dwell times. This gives us the arrival and dwell times of each user in each cell along a path.

5.2.2 Numerical Case Study: Senegal

In this section, we run the proposed imputation Algorithm 5 on real data, with different weight functions and fit different probability density functions, to identify the optimal model.

The 2014 D4D Challenge data sets

This study is based on the 2014 D4D Challenge data sets [4] provided by Orange Senegal.

- **D4D SET1 (Inter-site communications).** SET1 explains the macroscopic network behaviour. This part of the data set consists of aggregate call/SMS activities per hourly time slot and source-destination base stations. This data set includes the entire user base of the network.
- **D4D SET2 (User trajectories).** SET2 explains the microscopic, per-user view. This part of the data set consists of successive user activities with truncated timestamps (rounded to 10 minutes) and the identifier of the associated antenna. This data set encompasses 300,000 sampled users over a period of 2 weeks, and provides access to more fine-grained information on this user subset.

Inter-Antenna Road Paths

We imported the Senegalese road network from OpenStreetMap and added the location of the provided antenna sites into a spatial database. We precomputed the inter-antenna routes with their distances and temporal costs using the Dijkstra algorithm within the spatial database framework[2].Using the (about 2.5 million) inter-antenna routes, we computed their intersection with the Voronoi polygon geometries of the base station sites. Thus, we obtained the sequences of expected sites visited along all the inter-antenna routes .

Examples: Attractivity Metrics

Global Attractivity

Figure 5.2 shows a heatmap of global attractivities (expressed as a square root for visualization purposes) in Senegal for an example time slot (12:00-13:00). Note the low

[2] The computations were performed using the open source PostGIS and pgRouting packages

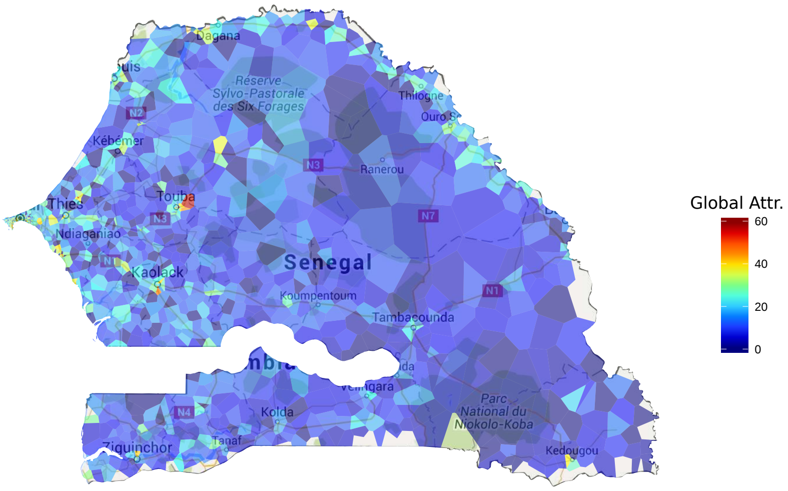


Figure 5.2: Voronoi representation of the Global (objective) Cell Attractivity (A_G) of sites in SET2, 12:00-13:00 time slot

attractivity of desert areas in contrast to the higher attractivity in the coastal regions, the capital Dakar peninsula and the mining regions in the south east.

Individual Attractivity

In Figure 5.3, we can see a single user's daytime attractivity metric overlaid on a map of Dakar. Note that there are three cells with particularly high attractivity (highlighted with white circles), which are likely to be the home and/or work locations.

Model Selection of Attractivity Weight Function

We now evaluate how to combine information from SET1 and SET2, which means comparing different functions to combine A_I and A_G.

Candidate Functions

We evaluated five functions for $F(A_G, A_I)$ within Algorithm 5 from Section 5.2.1. We chose them on the criteria of being simple, linear functions that include one or both attractivity metrics either linearly or squared:

- $F_1(A_G, A_I) := A_G$ as a baseline approach using only aggregate activity (i.e. global attractivity)

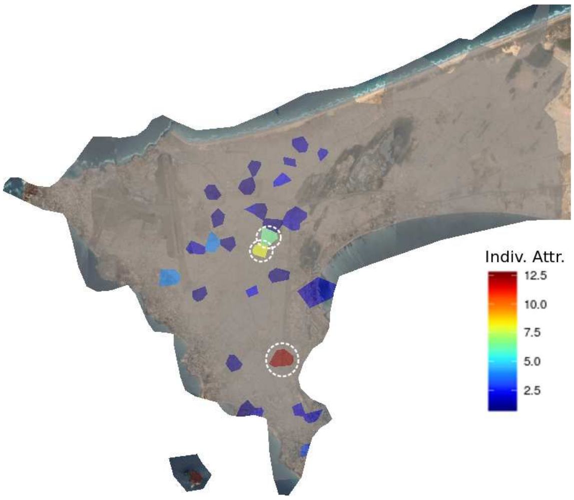


Figure 5.3: Daytime Individual (subjective) Cell Attractivity (A_I) of a single user from Dakar (favourite Voronoi cells)

- $F_2(A_G, A_I) := A_I$ as a baseline approach using only the per-user activities (i.e. individual attractivity)
- $F_3(A_G, A_I) := A_G A_I$ as a multiplicative model of both metrics (thus giving equal impact to both global and subjective attractivity)
- $F_4(A_G, A_I) := A_G^2 A_I$ similar to F_3 but giving more impact to global attractivity
- $F_5(A_G, A_I) := A_G A_I^2$ similar to F_3 but giving more impact to individual attractivity.

Precomputation

We ran our model on one month's data from SET2, omitting inter-activity times longer than 24h, in order to avoid large spatio-temporal "gaps" that are hard to impute in absence of additional data. We imputed the intermediate locations of users, and substituted the dwell time weight functions F_1 through F_5 for each visited site and arrival time. This means that the sites visited are identical for the different models, but that the dwell times are different, as the temporal split of the time between activity differs, depending on the attractivity weight function F used.

Dwell Time Percentiles

We list the ranges of the resulting dwell time percentiles in Table 5.1. These ranges represent the minimum and maximum percentile value over all antennas and the entire day, i.e. 24 hourly time slots. E.g., $F1 = A_G$ produces a median (percentile 50) dwell time over all antennas ranging from 300 to 1,300 seconds depending on the time of day. We can see the impact of A_I (assumption 2) on the longest dwell times (percentile 95). We reject $F_1 = (A_G)$ on the basis that the resulting dwell times are too evenly distributed and do not represent the heavy-tailed distribution that we need to model nightly stays, as only $< 5\%$ of dwell times are longer than 19,000s ($\approx$ 5 hours and 20 minutes). Based on these results, we continue our analysis on F_2, F_3, F_4 and F_5.

P	$F_1 = A_G$	$F_2 = A_I$	$F_3 = A_G A_I$	$F_4 = A_G^2 A_I$	$F_5 = A_G A_I^2$
50	[300;1,300]	[200;1,000]	[160;820]	[120;600]	[50;210]
95	[6,000;19,000]	[5700;27,800]	[5500;30,000]	[5400;32,000]	[5500;35,590]

Table 5.1: Rounded dwell time percentiles (in seconds) over all sites resulting from different attractivity weight functions

Cell Dwell Time Probability Density Function

To further evaluate which of these functions is the best formula for our model, we fit different probability density functions (PDFs) to the data generated by the four remaining candidate functions. The goal is to find a weight function F that produces dwell times that are distributed according to one of the PDFs that have been observed in real data. These are heavy-tailed distributions, as there are typically short transitionary dwell times from moving users, and longer dwell times from stationary users. In the relevant literature, among the most commonly used distributions are the exponential [109, 118], lognormal [117] and power-law [117,119] distributions.

Sampling

We uniformly sampled 500 tuples of $[hour, site]$ with $hour \in [0, 23]$ and $site \in [1;\ 1,666]$. For these spatio-temporal samples, we evaluated how many dwell time samples were available, each one corresponding to a single site visit of a user travelling between activities. We rejected tuples for which there were less than 100 user dwell time datapoints, to ensure

Model	F2		F3		F4	
PDF	$\bar{p}_{A_I}$	$\%^{A_I}_{>0.05}$	$\bar{p}_{A_G A_I}$	$\%^{A_G A_I}_{>0.05}$	$\bar{p}_{A_G^2 A_I}$	$\%^{A_G^2 A_I}_{>0.05}$
Power-Lognormal	0.585	90.4	**0.592**	90.0	0.572	89.2
Johnson SU	0.511	87.6	0.529	88.6	0.518	88.4
Fisk	0.506	91.0	0.499	88.4	0.447	84.0
Log-Normal	0.496	87.6	0.516	87.8	0.508	87.2
Johnson SB	0.484	87.0	0.497	85.6	0.485	82.8
Burr	0.486	88.0	0.470	86.6	0.440	81.2
Mielke's Beta-Kappa	0.455	85.4	0.459	87.2	0.422	83.0
Lomax	0.436	85.6	0.444	87.8	0.398	84.4
Generalized Pareto	0.440	85.6	0.439	88.0	0.409	86.0
Beta Prime	0.442	86.0	0.431	83.6	0.383	78.8

Table 5.2: Comparison of mean p-values for the 10 best-fitting probability density functions and attractivity weight functions calculated on 500 $[site, hour]$ tuples

the significance of the statistical test. We compared the P-Values of the four different attractivity weight functions F_2 through F_5 and of different probability distributions using a Kolmogorov-Smirnov-Test.

Distribution Fitting

In Table 5.2, we summarize the results of this statistical test on both data sets, listing the 10 best-fitting distributions. We omitted $F_5 = A_G A_I^2$ from the table since the best mean p-value we could obtain was < 0.004. Thus, we also reject it for our model. As we can see, the **power-lognormal** distribution performed best in combination with the $\mathbf{F_3 = A_G A_I}$ function at an average p-value of 0.592, with 90% of sites and hours tested passing the KS-Test. The power-lognormal distribution is a generalization of the lognormal distribution which has been used in other studies to model dwell times obtained from real data, both as single distributions [117] and in mixtures of multiple lognormals [120]. As a heavy-tailed distribution, it captures both long- and short-term stays within cells, as users are either on the move or stationary when performing an activity. This supports the adequacy of our model for dwell time estimation.

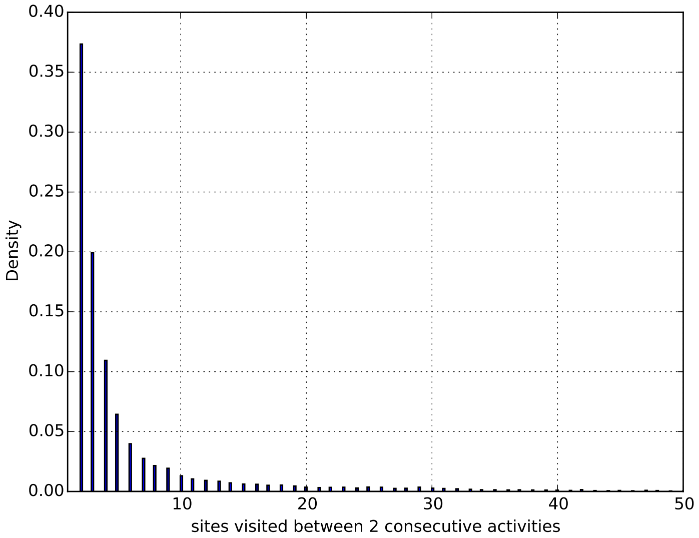


Figure 5.4: The estimated number of sites visited in SET2 between user activities

5.2.3 Results and preliminary validation

Having identified the model as the attractivity weight function $\mathbf{F} = \mathbf{A_G A_I}$ and the **power-lognormal distribution**, we now present the results based on using this model. As we do not have the matching precise ground truth data, we will consider different other data sets and studies and compare our results as a preliminary validation.

Aggregate Mobility

We can see in Figure 5.4 that most inter-activity trips correspond to short trips of fewer than 10 sites visited. The distribution follows an exponential decay, as is to be expected if we assume that the number of sites visited is correlated to the distance travelled, as this occurs similarly in human mobility, e.g. in taxis, as studied by Liang et al. [119]. It also follows the same pattern as the number of distinct cells visited per day in the Orange France 3G data set examined by Hess et al. [103]; due to the symmetric nature of most commuting trips, both statistics are comparable.

Figure 5.5 represents the mean number of locations (base stations) visited per hour, which we compute as $3600 \times mean(DwellTime(t))^{-1}$. The aspect of the curve corresponds to other studies of human mobility, e.g. the findings of Demissie et al. concerning handover

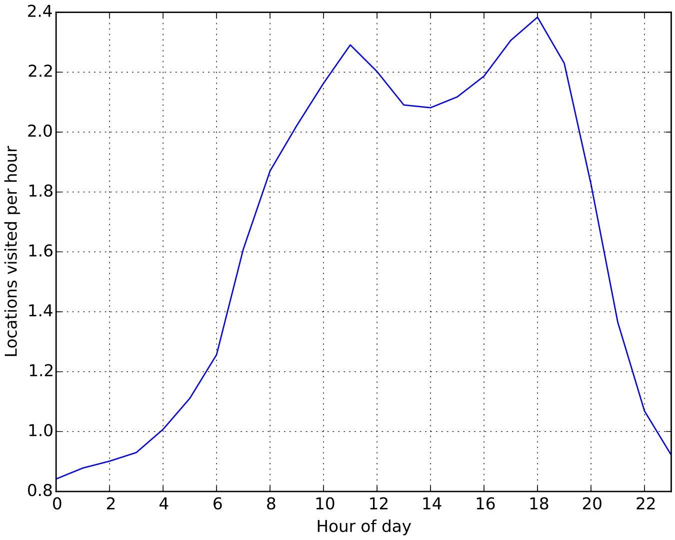


Figure 5.5: The estimated mean number of base stations visited per hour

and mobility data from Lisbon [121], and those from Sagl et al. on data from Amsterdam [122]. There is a significant slow-down of mobility at night (between 0:00 and 6:00), and an increase in mobility at daytime, with two peaks at 11:00 and 19:00, indicating the rush hour commuter traffic and the effect of increased mobility within cities. The order of magnitude of the mean number of locations in [0.8; 2.4] is sensible in comparison to the findings of Hess et al. [103], with values within [1.05; 1.5], considering that we get all intermediate locations instead of only connected locations.

Dwell Times

Figure 5.6 displays the quantiles 0.5, 0.8, 0.9 and 0.95 of the estimated dwell times for all antennas over an entire day. As expected, we can see that dwell times are much longer at night, while between 10:00 and 19:00, they are shorter due to the increased daytime mobility.

In Figure 5.7, we can see the mean dwell times between 12:00 and 13:00 on a country-wide scale. We omitted the bottom 10% of sites ordered by the estimated number of user movements in the site to improve readability. The estimated mobility is concentrated along borders and main roads. In the west, there is significant mobility in the coastal regions, but there is also arterial and rural mobility, of which we see less in the east.

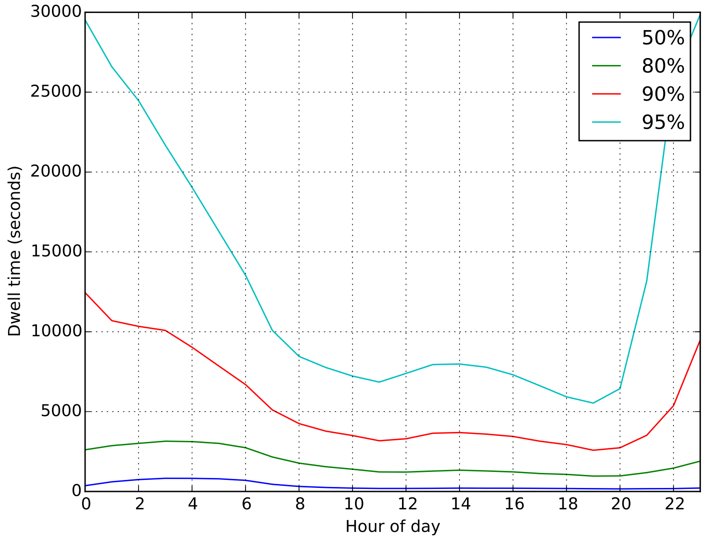


Figure 5.6: Quantiles of dwell times generated for all antennas by hour of day

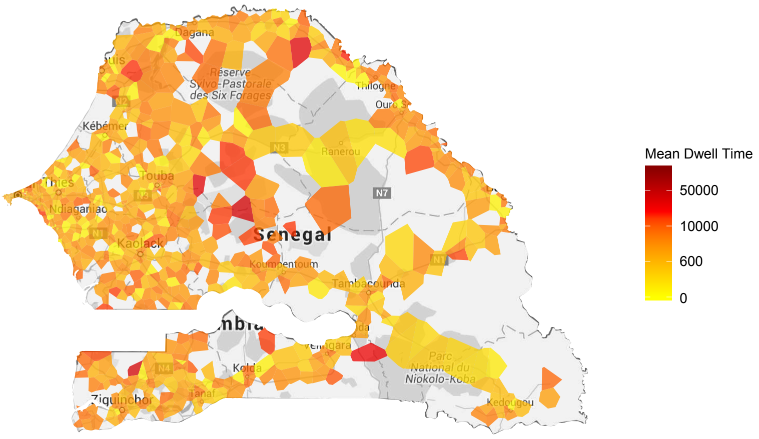


Figure 5.7: Heatmap of estimated dwell time means given arrival between 12:00 and 13:00

Figure 5.8 shows the same type of map, but for the midnight timeslot, corresponding to dwell times given an arrival between 0:00 and 1:00. Mean dwell times are higher than at noon (on Fig.5.7) because users arrive at their home location and stay there for the night. Most cities and suburban regions exhibit long dwell times.

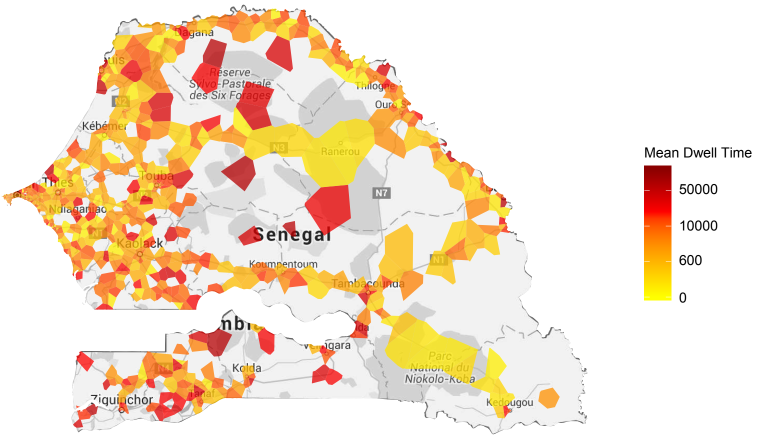


Figure 5.8: Heatmap of estimated dwell time means given arrival between 0:00 and 1:00

Comparison to real-world dwell time data

We compared the resulting dwell times to signalling data from China Mobile's GSM and UMTS networks as published by Zhou et al. [117]. Their data set encompasses 15,000 base stations covering 3,000 km^2. We extracted data from their dwell time distribution plots, which were aggregated by day- and nighttime between 9:00-17:00 and 22:00-6:00, respectively. As their PDF parameters were fitted on the majority of samples in the short dwell time regions, we created a logarithmic-scale regression model from the data points we extracted to capture the long dwell times more adequately, as formulated in Table 5.3. The results indicate that there is a significant shift in cumulative distribution in the shorter (0 to 10 minute) dwell time region. This is due to the ping-pong effect of terminals changing between cells rapidly in presence of low differences in signal strength (only encountered in real data), and the fact that inter-activity times are truncated to 10 minute precision in the D4D data set (cf. Section 5.2.2). This means that our inter-activity times are at least 10 minutes, and thus exclude accurate estimation of the shortest dwell

	Day	Night
Intercept	-0.249	-0.565
t	-1.720	-1.296

Table 5.3: Logarithmic scale regression for day- and nighttime dwell times in the China Mobile data [117]

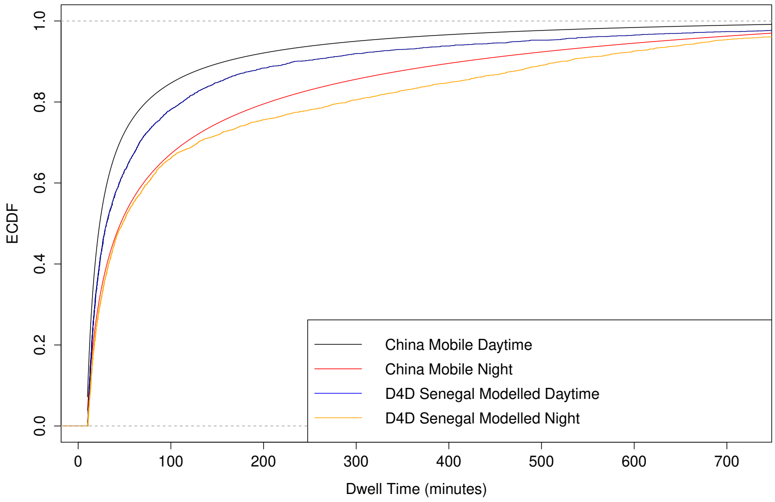


Figure 5.9: Comparison between day- and nighttime dwell time ECDFs (dwell times > 10 minutes), China (real data) and Senegal (modelled)

time percentiles. We plotted the Empirical Cumulative Distribution Function (ECDF) of dwell times longer than 10 minutes in Fig. 5.9. As in [117], we aggregated dwell times from our model by day- and nighttime between 9:00-17:00 and 22:00-6:00, respectively. We can see that our modelled dwell times closely reflect those encountered in the real world, both at day- and nighttime, within a distance of 5% cumulative probability.

5.2.4 Discussion

In this study we have presented a model of mobile network cell dwell times exploiting information from two anonymized CDR data sets of the D4D Challenge. We have introduced a novel way of estimating cell dwell times based on cell attractivity factors, both at single-user and network scale. This enables the creation of cell dwell time distributions using trajectory data from only a small part of the population. We identified the product of individual and population-wide activity metrics to be the best weight factor for temporal imputation. The generated data could be fit best by the power-lognormal distribution, for which the distributions passed the KS-Test in over 90% of cases. The fact that this distribution fits our data best is in strong agreement with results on real dwell time distributions.

Currently, a direct validation of our model is not possible due to the lack of ground truth data that we would require, i.e. handover data corresponding to a CDR data set, in this case Senegal. This could only be achieved with the direct support of a mobile network operator. It is for this reason that we opted to validate our results with dwell time data and related statistics from other countries: Most importantly, we compared our resulting dwell times to the study by Zhou et al. [117] on data from China Mobile (section 5.2.3). We observed that, on an aggregate level, the cumulative dwell time distributions generated by our model at day- and nighttime only differed from those in the China Mobile data set by less than 5% for dwell times longer than 10 minutes. Also, we have shown that our results agree in the number of distinct cells visited per day and the corresponding hourly mean in the Orange France 3G data set as examined by Hess et al. [103] (section 5.2.3). Additionally, we compared the mobility patterns we observed in our modelled data to the findings of Demissie et al. regarding data from Lisbon [121], and findings from Sagl et al. on data from Amsterdam [122], and found that our mobility followed similar within-day trends.For some applications, our model eliminates the necessity of handling and mining large CDR data sets, and replaces it by the use of a simple, parameterized model, e.g. for modelling population mobility in epidemiology. Another advantage is that the model's input CDR data can be generated effortlessly by mobile network operators. If they were to make use of a model like the one we proposed, they could make the population dynamics inside their coverage area available for a wide range of studies, to the benefit of the research community.

5.3 Constraining Demand Estimation Models with Signaling Data

One critical step in demand estimation is creating the seed matrix, i.e. the initialization matrix (or matrices) serving as a starting point for the optimization problem. In order to facilitate the process of generating the seed matrix, we evaluated how mobile network signaling data can constrain this problem. In [Cantelmo et al., 2017], we used signaling data from POST Luxembourg, i.e. hourly aggregated handovers between base stations. We took all radio access technologies into consideration, and span a border around Luxembourg City to evaluate the amount of flows entering and exiting the city. Fig. 5.10 shows the hourly proportions of flows in both directions.

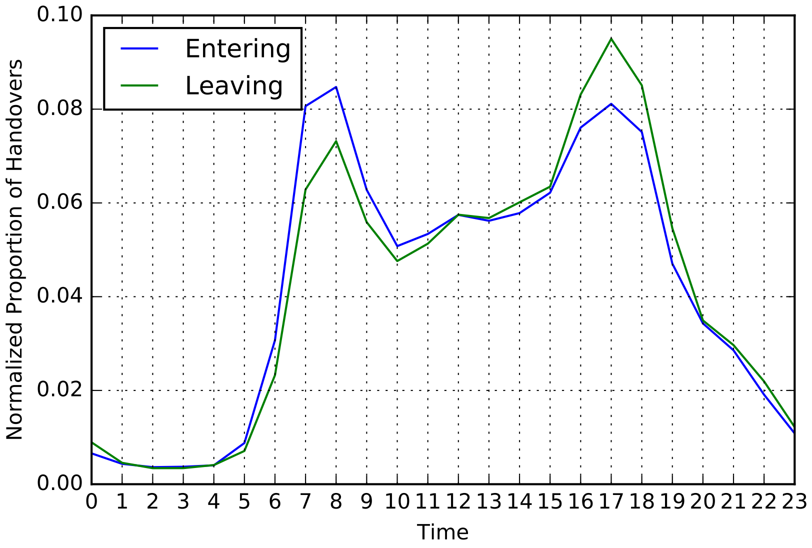


Figure 5.10: Proportions of handovers entering and leaving the Luxembourg City area.

Fig. 5.11 shows the difference between both directions, clearly highlighting both peak commuting hours. These metrics can be used for attraction and generation in the demand generation step, and support creating informative initial Origin-Destination (O-D) matrices for subsequent optimization.

The results obtained in [Cantelmo et al., 2017] showed that using mobile network handovers as an additional constraint allowed for a much faster convergence of the O-D estimation, supporting our claim that mobile network data is a valuable source for demand estimation and activity-based modelling.

5.3.1 Application Case Study: Multimodal Trip Planner

In [Faye et al., 2017], we show how a calibrated demand can be used in an Advanced Traveller Information System (ATIS) and simulation setting. In particular, we setup a multimodal trip planner based on the OpenTripPlanner framework. As shown in Figure 5.12, the platform is composed of three main components. At the core, a multimodal route planner evaluates multiple data sources and computes travel itineraries through a graph-based routing approach. The trip planning service, which is publicly available online, is directly connected to two other modules. One module is a calibrated multimodal network model of Luxembourg in PTV VISUM, which makes it possible to estimate real-

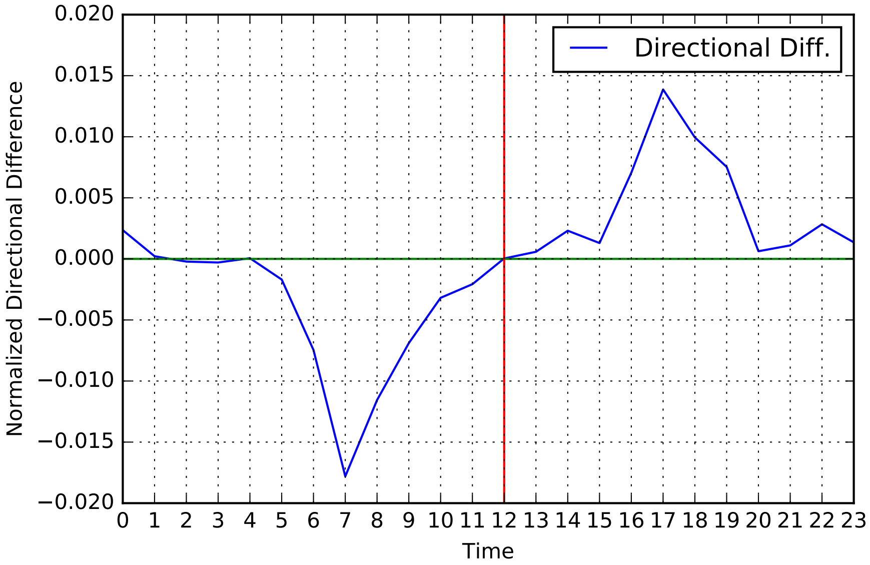


Figure 5.11: Directional Difference between handovers entering and leaving the Luxembourg City area.

istic traffic demand or to generate high-demand scenarios on-the-fly. The link travel times computed in VISUM are then used to update the routing graph of OpenTripPlanner. The other module is a mobile application that allows users to anonymously estimate different aspects of their daily mobility and to personalise the route planner settings and its interface.

The system considers private and public transportation, including the Luxembourg bike-sharing system along with rental bike availability. The main strength of the system lies in the possibility to integrate the simulated reaction of the network to perturbations (as computed in VISUM) in the routing graph. This leads to adapted route recommendations, even in the absence of large sets of FCD. This platform is available online[3] and also offers multimodal isochrone maps for different modes. To summarize, in absence of large-scale FCD, calibrating a realistic demand with the help of mobile phone data and integrating this demand model in a simulator can enable a realistic evaluation of various traffic anomalies. The simulation can then serve as an input for ATIS and estimate e.g. the anticipated impact of road construction sites and provide users with better path alternatives, avoiding the concerned road segments.

[3] http://otp.mamba-project.lu

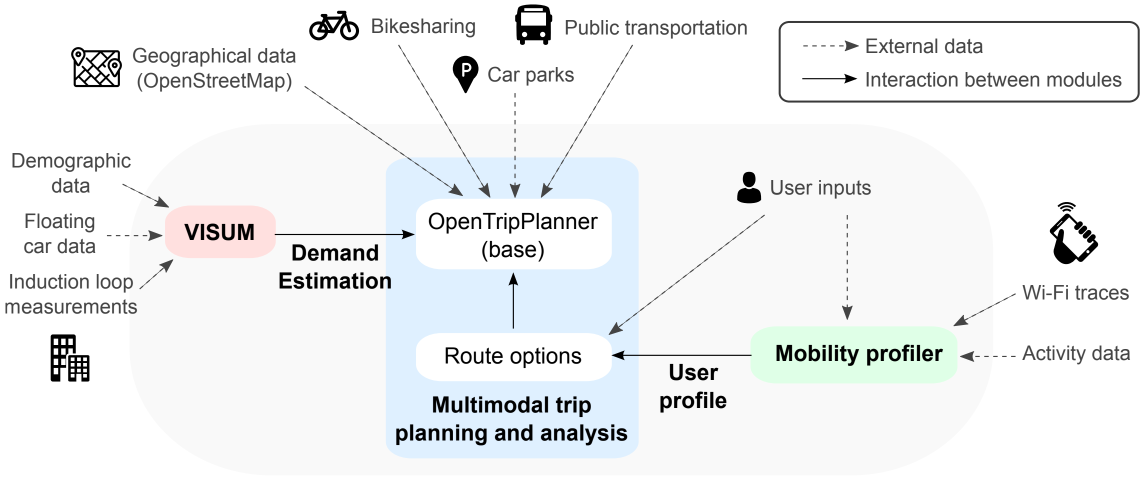


Figure 5.12: Trip Planner Overview

To summarize, the presented system leverages conventional and mobile network data to produce realistic travel time estimates and useful route recommendations using simulation.

Chapter 6

Traffic State Modeling using Mobile Network Data

Estimating traffic states from mobile network data –especially in urban environments– is a new and promising research topic. It involves identifying links between traffic flow theory and the observations in the mobile network. In this chapter, we want to define how urban traffic states can be estimated using mobile network signaling data, in particular aggregated handover counts.

6.1 State of the Art

Mobile network handovers exist in two varieties: on one hand, passive handovers of phones that are currently not in an active phone call or data session; their location is known to the network at *Location or Region Area* (LA/RA) level, encompassing potentially hundreds of mobile cells. On the other hand, active handovers of phones in a connected state provide information of the exact currently associated mobile cell.

A lot of research has been focusing on passive handovers, i.e. coarse-grained *Location Area Code* (LAC) updates which can be useful in predicting highway travel times [6]. The main work in this area is a study by Janecek et al. [90], who combine location updates to the handovers of active calls along a specific highway in Austria. They study the rate of LAC updates from idle mobile phones and augment this knowledge with the rate of active connection handovers to clearly identify and precisely locate the source of congestion. However, this methodology is valid for highways only and it is difficult to extract the required data for larger areas.

In general, passive handovers (LAC updates) cannot be used for state estimation in urban environments. They are useful for long-range travels, as studied by Hui et al. [123]. In this work we want to investigate how mobile network data can be used for estimating congestion within cities, by using only aggregated active connections. For these connections, the precise cell rather than a large location area is known, leading to a much higher spatial resolution even when computing aggregate statistics. In this vein, Bar-Gera [124] ran a study on using active connection handovers to predict freeway travel times, using probe mobile phones to record both the handover events and travel times and comparing the measurements to loop detector data. Again, this study focused on highways and not on urban settings. Another limitation of cellular datasets for traffic flow estimation is that they include mobile and static nodes. To overcome this limitation, Caceres et al. [125] proposed a set of models to infer the volume of vehicles from the cellular data by calibrating them with data collected by loop detectors. On average their best model achieves an absolute relative error of less than 20% for highway scenarios. They conclude that cellular data can be used as a complement to traditional fixed sensors to enhance the available information for mobility monitoring.

Generally speaking, there is little research regarding traffic states in urban environments, as Naboulsi et al. identify in their survey [6]. The main study in the urban traffic area was done by Calabrese et al. [62], who performed analyses of the Telecom Italia dataset for the city of Rome, in particular Erlang data (a unitless metric of the intensity of mobile network usage) alongside taxi and bus data. This allowed them to build a platform to estimate what they call the *pulse* of a city, and to compare the availability of public transportation to their estimated population location density.

The correlation between the road traffic state and the observed reaction of the mobile network is an interdisciplinary topic, connecting transportation and telematics. It is therefore sensible to rely on concepts from traffic flow theory such as the Macroscopic Fundamental Diagram (MFD), which describes the traffic profile of an urban area from a macroscopic, aggregated perspective. MFDs are synthetic but powerful metrics that quantify and explain the interaction between road capacity, travel and driving behavior-related parameters such as routing/rerouting, as well as characteristic vehicle speeds and car following behavior. It postulates that if a sufficiently large amount of data about traffic states in a network is collected, and the (sub) road network topology has a sufficient level of regularity in terms of route flow distribution, then state variables such as vehicle density and the total network throughput are clearly related by a concave function. This function expresses the transition between uncongested conditions to congested states, characterized in urban systems by frequent conditions of queue blocking and gridlock phenomena. Theoretical and empirical studies contributing to gain insight into the properties of MFD focused on deriving relations starting from analytical and simulation-based Dynamic Traffic Assignment theory [23], on assessing the impact of traffic control [126], and on capturing hysteresis phenomena in congested networks [127].

6.2 Traffic State Estimation from Signaling Data

In this study, we want to estimate traffic flows from a 4G mobile network dataset. The dataset is composed of two components. The first is the position of LTE base stations (eNodeBs) and the corresponding cell identifiers hosted on this base station. The second is the number of handovers between any given cell pair per hour. In the remainder of this section we will refer to the handovers within a set of cells as *internal flows (i)* of that set, and to the handover count leaving a set of cells as its *exiting flows (o)*. Both metrics are scaled into $[0, 1]$ with respect to their daily maxima. We will also refer to the *traffic state t* as the space-mean of the ratio between actual velocity and the legal speed limit $(\overline{v \div v_{max}})$. Further details on the different datasets used will be provided in Sections 6.2.3 and 6.2.4.

6.2.1 Methodology

We want to establish a model in the form of $v = q \div k$, i.e. the fundamental flow-density relationship for partitions of the road network, in analogy to the concept of MFDs. Since in mobile networks the phone's precise serving cell is only known during an active data or call connection, we cannot access the density of mobile phones directly (as the majority of them typically are in a passive, disconnected state). Thus, we propose a three-stage approach:

First, we partition the road network in areas that are large enough to capture the traffic dynamics of MFDs. Next, we model each partition's density using handovers within and from the partition. Finally, we use linear regression to estimate the traffic state from exiting flows and approximated density, thus optimizing the regression coefficients globally

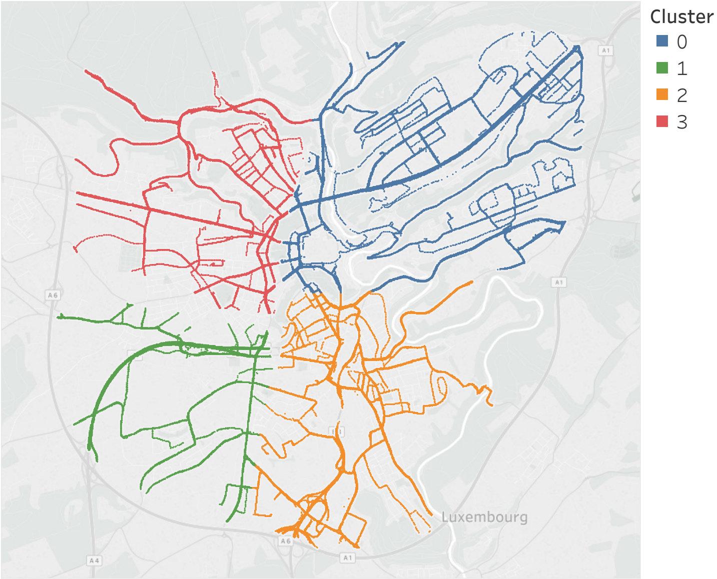


Figure 6.1: Luxembourg City road network partitions used in both simulation and real data study

for all time intervals and partitions.

Stage 1: Network Partitioning

In this chapter, we will focus on theoretically and empirically studying the traffic and mobile networks of Luxembourg City[1]. Fig. 6.1 shows the partitioning we opted for, which we will use both in the simulation and real-data studies. The study area covers approximately 45 km^2. According to Daganzo and Geroliminis [22], MFDs emerge in areas larger than 10km^2. Thus, we opted for 4 partitions, representing the main geographical zones of Luxembourg City, i.e. physically separated plateaux, independent from the number of flows. Note, however, that road network partitioning can also be done algorithmically and depending on the flows, e.g. using the normalized cuts [128] or spectral clustering algorithms [Derrmann et al., 2017a], or be based on data concerning mobile phone calls [13, 129].

[1] Center coordinates: 49.611634, 6.129451

6.2.2 Stage 2: Traffic state and density models

We want to define density and flow proxy functions to predict the traffic state in analogy to the fundamental equation of traffic flow ($v = q \div k$). The goal is to use each partition's scaled internal (i) and exiting flows (o) and a density proxy function characteristic of it – referred to as $k(i, o)$ below – so as to obtain an estimate of the current traffic state (t) within the partition. We express this idea using linear regression with a logarithmic transformation:

$$log(t) \sim a\ log(o) + b\ log(k(i, o)) + c \tag{6.1}$$

$$t \sim \frac{o^a}{k(i, o)^{-b}}\ exp(c) \tag{6.2}$$

In Eq. 6.2, we require a density modeling function $k(i, o)$ based on the ratio between the scaled inner and exiting flows ($i,\ o \in [0, 1]$). Following some preliminary results with a simple quadratic relationship, we propose to model the relationship using a more expressive polynomial with interaction, where the degrees (p_i, p_o, p_{ix}, p_{ox}) and coefficients (c_i, c_x, c_o) are the parameters characteristic of each partition:

$$k(i, o) := c_i\ i^{p_i} + c_x\ i^{p_{ix}}\ o^{p_{ox}} + c_o\ o^{p_o} \tag{6.3}$$

This model can then be easily interpreted by visually inspecting the surface of the polynomial in the three dimensions (i,o and their resulting modelled density $k(i, o)$).

On the level above the density proxy function, we then have three global parameters that are shared between all the partitions ($[a, b, c]$) and need to be estimated to link the flow and density proxy functions into a traffic state:

$$\tilde{q} := o^a \tag{6.4}$$

$$\tilde{k} := k(i, o)^{-b} \tag{6.5}$$

The unit of k is $(veh.\ m^{-1})^{-\frac{a}{b}}$. An approximation of the space-mean density $\overline{\rho}$ with respect to the space-mean speed limit velocity $\overline{v_{limit}}$ is given by:

$$\overline{\rho} = \tilde{k}(i, o)\ o^{(1-a)}\ \overline{v_{limit}}^{-1} \tag{6.6}$$

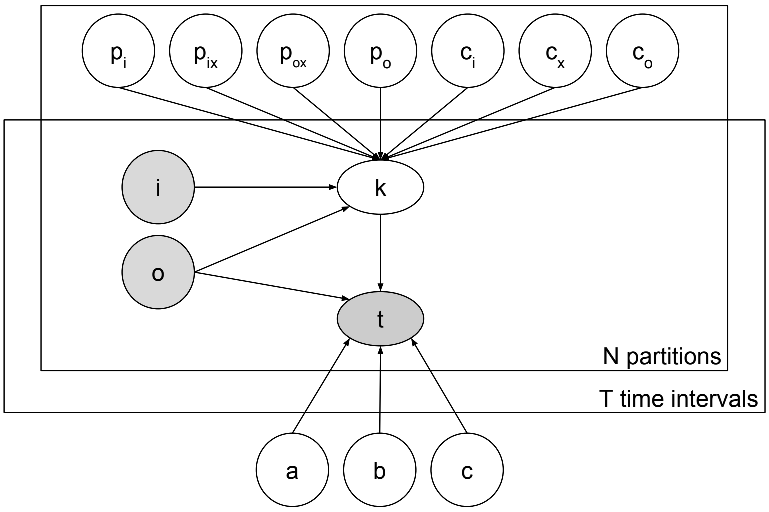


Figure 6.2: Graphical representation of the model: p and c are characteristic of each partition, $[a, b, c]$ are common parameters.

Fig. 6.2 gives an overview of our full model in graphical form. The shaded circles represent observed variables (in the training set) and the unshaded circles are latent estimated parameters or dependent variables in the case of k. For each of the N partitions, we estimate the parameters of the polynomial density model k, i.e. the p and c parameters. Its inputs are the scaled internal and exiting handovers i and o for each interval. The density k is then used alongside o as input for the linear regression model, that is globally parameterized (i.e. for all partitions and across the entire time range T) by its coefficients a, b and c. Essentially, we approximate density, then use the flow-density relationship to estimate the traffic state.

Parameter Estimation

In order to estimate the four degree parameters $(p_i, p_{ix}, p_{ox}, p_o)$ and three coefficients (c_i, c_x, c_o) in the density polynomial of each partition (Eq. 6.2.2), we implemented a hill-climbing optimizer.

Fig. 6.3 shows how the data set is used in this approach. We start from a random vector of density polynomial parameters in $[0, 2]$. During each iteration, we update the density parameter by adding a random offset sampled from $Uni[-0.01, 0.01]$ to a single parameter. Next, we run linear regression on our model (Eq. 6.2) and evaluate the resulting

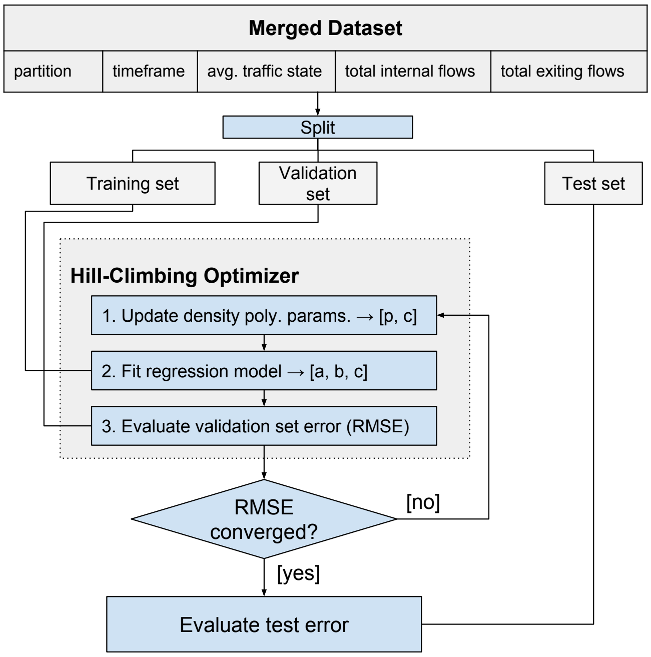


Figure 6.3: Parameter estimation: Hill-climbing optimizer

Root Mean Square Error (RMSE) of the validation set. The goal is to find the density proxy polynomials of each partition that allow the best regression performance. We accept parameter updates that lead to a lowering in validation RMSE, and that yield $a > 0$ and $b < 0$ in the linear regression step. The latter conditions are to assert that the density model can be interpreted as intended, i.e. $k(i, o)$ is a directly proportional proxy of the true density and $\tilde{v} = \tilde{q} \div \tilde{k}$ is respected.

Validation Techniques

In order to validate the model, we evaluate its predictive power on test data sets. We use the same methodology – i.e. partitioning and prediction model – for both the simulation and real-world studies so as to be able to compare them, and to be able to quantify the impact and limitations of the simulation.

6.2.3 Simulation Study

The LTE network configuration consists in a mapping of 113 eNodeBs (LTE base stations) to the simulation coordinates. The original coordinates of the eNodeBs were

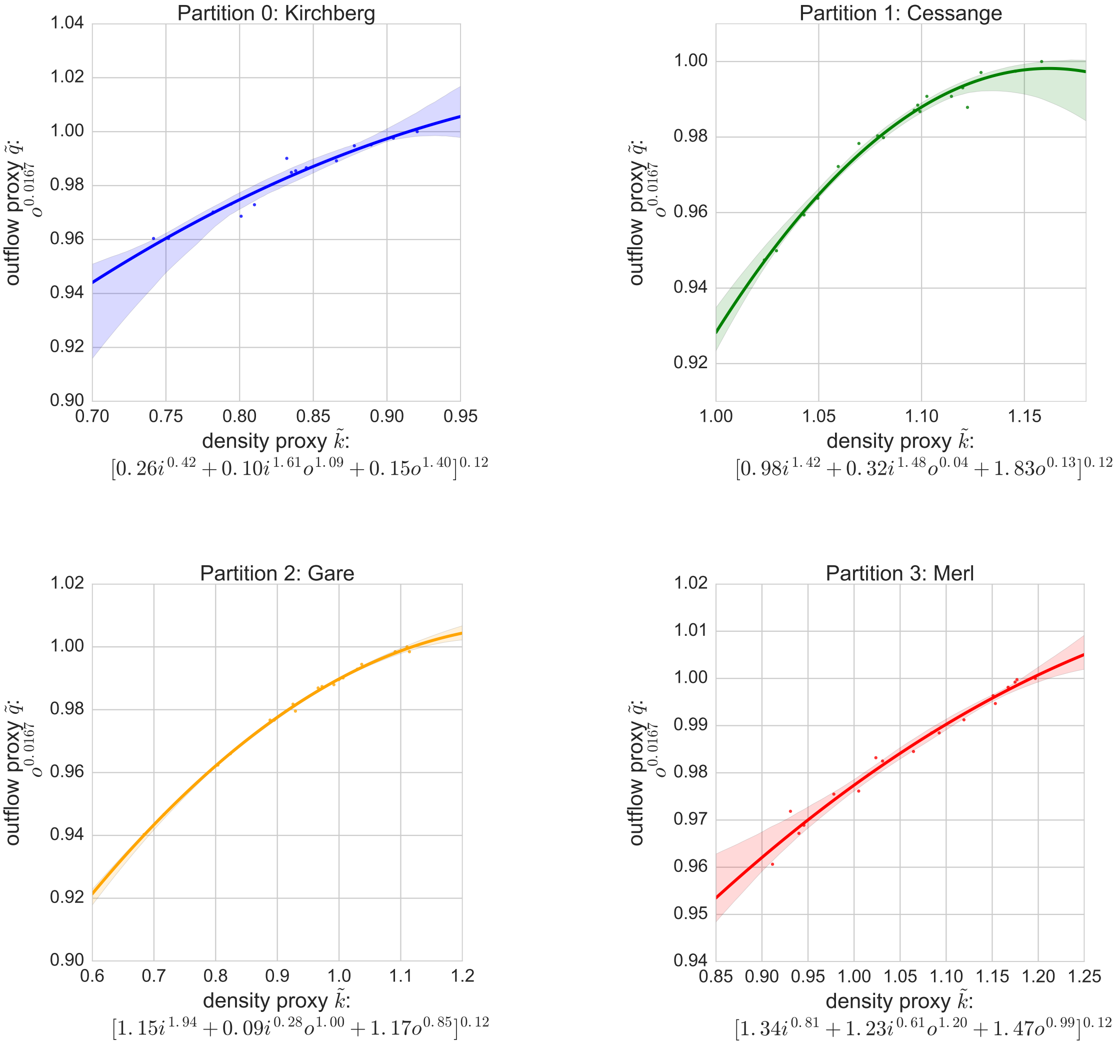


Figure 6.4: Simulation Study Mobile Network Macroscopic Fundamental Diagrams: Flow-density relationships by partition

provided by POST Luxembourg. Note that each eNodeB hosts multiple cells, but that the simulation does not account for the precise associated cell. The simulation framework, LuST-LTE, is published in [Derrmann et al., 2016a].

The simulation scenario we base our study on is the LuST scenario by Codecà et al. [77] for the microscopic traffic simulator SUMO [33]. The scenario provides 24 hours of mobility consisting of almost 300000 vehicle trips in a wider area around Luxembourg City (155 km^2). As we are studying urban environments only, we limit our study to the inner city (within the highway ring), approximately 50km^2.

Artificial Datasets

From the SUMO simulator, we obtain vehicle positions and velocities, i.e. simulated floating car data. This information is augmented with the currently connected cell, allowing us to compute the *Space-Mean Traffic State* $t = (\overline{v \div v_{max}})$ within the coverage area of a set of mobile base stations.

From OmNET++ and SimuLTE, we extract the number of handovers between cell pairs observed. Since we know the mapping between base stations and road partitions, we can compute the internal flows (i) and exiting flows (o).

In order to construct the data set, we ran the scenario with 50% re-routing probability of vehicles and 300 second re-routing interval, which were the most realistic parameters according to the validation by Codecà et al. [67]. The penetration rate of vehicles in active calls was defined as 1%, which is in line with a previous study by Caceres et al. [125]. The data set split was defined as a 50-50 split of the data, where validation and training sets both make up 25% and the test data is 50%. We opted for this split because otherwise modifying the demand and running an additional simulation day would have made the prediction error directly dependent on the degree of modification of the demand distribution.

As temporal scale, we chose 1 hour, yielding sufficient number of training and test data points (48 of each, i.e. 12 hours with 4 partitions), and matching the real data studies that we perform in the second stage.

Results: Mobile Network MFD Proxy

As described in Section 6.2.2, we estimated each partition's density proxy polynomial functions and the regression coefficients jointly using a hill-climbing optimizer.

Figure 6.4 shows the flow-density relationship resulting from the parameter estimation (as described in Sec. 6.2.2). We can see that Partitions 0, 1 and 2 show a tendency of saturation, and similar profiles in general. The resulting density proxy polynomials, however, differ strongly between the partitions, meaning that different ratios of internal-to-exiting handovers are characteristic of their traffic state profiles. The MFD of Partition 3 on the other hand, exhibits a quasi-linear flow-density relationship, indicating that this partition does likely not reach critical capacity and thus there is no reduction in flows caused by congestion. Thus, we do not observe the descending phase of the flow-density relationship, as we do for the other partitions. Overall, we do not observe the very harsh congestion MFD profiles that would be produced by grid-lock phenomena, but this is not the case in Luxembourg City, and is in line with other real-world results from other cities. The fact that these smooth, low-variance profiles result from our methodology is a first encouraging result, as they match the expected MFD shapes.

Results: Prediction

Fig. 6.5 shows the model predictions on the simulated data. On the y-axis, we see the actual mean traffic states of a partition during a 1-hour window, as computed by SUMO. The x-axis represents the predictions computed by the model using the simulated LTE signaling data, namely the internal and exiting flows (i) and (o) and the derived density proxy (k). The blue line shows the trend between both measures, which should ideally coincide with the green identity line. Since trend and identity lines are close, and the variance (error) appears to be stable across the range of true traffic states, we can conclude that the model fits the data reasonably well. The *Mean Absolute Percentage Error* (MAPE) is 10.2%, which is an encouraging result given the simplicity of the proposed model and the low amount of training data.

Limitations

There are several limitations in the simulation study.

On the road network side, there are only vehicles, no pedestrians. There are also no stationary users, that might impact mobile network handovers by moving minimal distances and triggering ping-pong handovers.

On the mobile network side, there is the inherent error of our model of purely SNR-based handovers versus real handovers, that are much more complex in nature. Further, we only simulate the LTE network connectivity, thus omitting the other radio access technologies, which influence handover behavior as well, e.g. through interference, and intra-RAN handovers. The fact that we only associate vehicles to eNodeBs, not cells, leads to additional error. Most importantly, we have a static penetration rate of 1% of vehicles in an active connection. While this is a realistic percentage on average, it is dynamic in reality in the course of a day as described by Caceres et al. in [125].

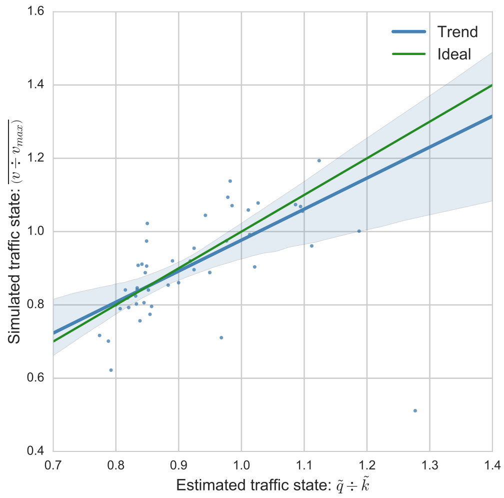


Figure 6.5: Simulated mobile data-based traffic state predictions vs ground-truth simulated floating car data

However, for the simulation study, we are interested in the general feasibility of handover-based traffic state estimation using our methodology, and not as much in the precise attainable error. Thus the limitations above should be considered but not overvalued.Having shown the performance of the model on simulated data, we will now evaluate it using real data to show its performance without the simulation limitations.

6.2.4 Real Data Study

Ground Truth Data Set: Floating-Car Data

As ground truth data, we use *Floating Car Data* (FCD) that was made available for a whole week at the end of September 2016. This is a set of time-stamped location updates and travel speeds which were collected in the area of Luxembourg City, and consists of 600 trips and 220000 location updates. In particular, we are interested in *Traffic States*, i.e. the ratio between actually driven speeds and the speed limit ($v \div v_{max}$). Thus, we performed map-matching on the FCD to obtain the values of v_{max} for every location update.

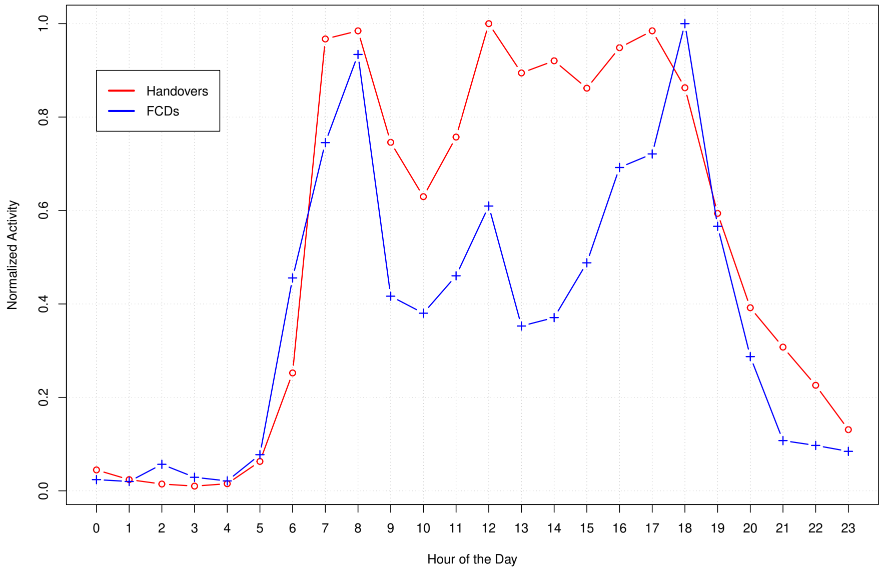


Figure 6.6: Normalized number of handovers observed in the study area vs. number of floating car data entries

Mobile Data Set: LTE Handovers

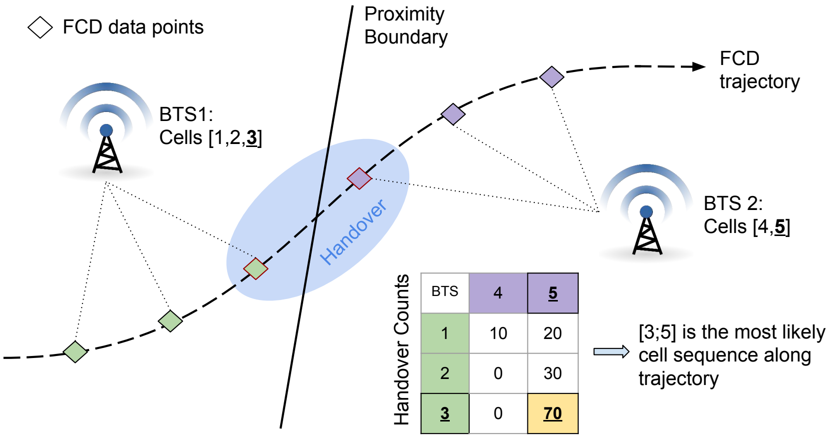
The mobile data set contains aggregate data from 436 LTE (4G) cells within the Luxembourg City. The data consists of the number of handovers between cell pairs per hour. The data was made available for the same time period as the FCD.

Fig. 6.6 shows that the number of handover and floating car observations correlates strongly, except for the off-peak daytime, when there are relatively more handovers, likely due to pedestrian movement and increased mobile phone usage. This correlation is a strong motivational aspect to our work, and we found similar correlations between mean travel speed and artificial handover counts in previous work [Derrmann et al., 2016b].

Mapping FCD to the Mobile Network

In order to enable the use of FCD for validation purposes, we need to map the most likely associated mobile network cell to each FCD data point.

Fig. 6.7 shows an example of the method we used: First, we can easily find the nearest *Base Station* (BTS) of each location by distance. In the example, that is BTS1 for the



BTS	4	5
1	10	20
2	0	30
3	0	70

Figure 6.7: Floating-Car Data and Mobile Network Mapping: Every vehicle position is matched with its most likely associated mobile cell (cf. Sec. 6.2.4).

first three floating car data points, and BTS2 for the next three. Usually, a BTS hosts a set of mobile network cells emitting into different directions, e.g. cell 1, 2 and 3 for the first part of the trajectory and cell 4 and 5 for the second part.

From an FCD trajectory, we can thus identify a sequence of these sets of potentially associated cells corresponding to the taken road path. Now, in order to identify the single, most likely visited cell sequence, we choose the most frequent cell transition during that day to be the likely cell pair visited. That way, we build a chain of visited cells over the entire trip. In the example above, the most likely cell transition (handover) is $3 \rightarrow 5$, because most handovers are between these two cells. Thus, we pick these two cells as the most likely occurred sequence.

Using this method, we get a single likely associated cell for each FCD entry, i.e. the cell that the driver's phone was most likely connected based on the current location. This allows to compute road traffic statistics relative to the connected cell.

In the last step, in order to compute the *Traffic State* variable t, i.e. the ratio between the actual observed link speeds and their respective speed limits, we perform map-matching of the Floating-Car Data entries to the *OpenStreetMap* (OSM) road network to obtain the speed limit at each entry.

Finally, the resulting merged data set contains the partition number, hour of day, internal and exiting flows ($[i, o]$) and the traffic state (t).

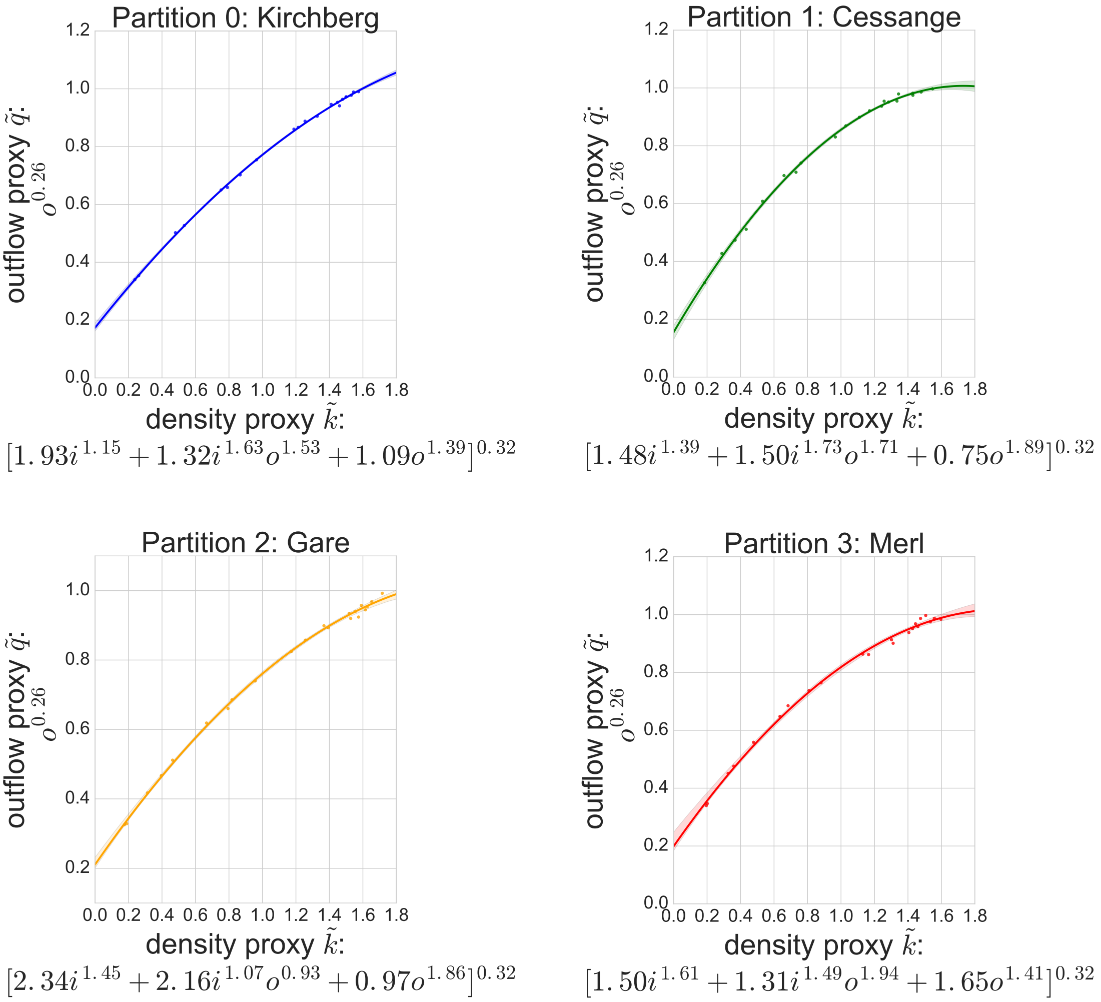


Figure 6.8: Real Data Study: Mobile Network Macroscopic Fundamental Diagrams: Flow-density relationships by partition

Results: Mobile Network MFD Proxy

Using the merged data set as described above, we trained our model on the data of Monday, Tuesday and Wednesday, validated it on Thursday and tested it on Friday data.

The linear regression model resulted in the following parameters:

$$\begin{cases} a = 0.26 \\ b = -0.32 \\ c = 0 \end{cases}$$

Since there is no intercept ($c = 0$), the scale factor is 1 (i.e. $exp(c)$ in Eq. 6.2). Thus the prediction equations of the space-mean traffic state t proportional to the space-mean velocity v yields the direct ratio of $\tilde{q} \div \tilde{k}$:

$$t \sim \frac{o^{0.26}}{k^{0.32}}$$

We plot the MFDs given by $\tilde{q} = o^{0.26}$ and $\tilde{k} = k^{0.32}$ in Fig. 6.8. Unlike in the simulation study, all the MFDs exhibit very low variance and as is expected, they all follow a concave shape as the outflow rate saturates at increasing density levels. Partition 1 (green plot) exhibits this behavior clearly.

The lower variance is likely due to the larger training set, as well as the larger number of handovers observed in the real world in comparison to the simulation, reducing the impact of noise. We can also see that there is no severe congestion in the network, caused by possible grid-lock phenomena, that would manifest itself in the descending phase of the MFD diagram. Instead, we only observe saturation of the network.

Results: Correlation and Prediction

Table 6.1 shows the Pearson correlation coefficient between the model's predictions and the real traffic states observed from the FCD. We observe moderate to strong correlations for all four partitions. The correlation values are slightly higher than for the simulation run, which we attribute primarily to the larger aggregation samples and time windows. They highlight the information content of the mobile network MFDs, and the fact that all four partitions could reflect the underlying traffic states well. As in the simulation study, the more heterogeneous Partition 0 shows the weakest fit. With a more adequate partitioning method, this could potentially remedied, as presented by Ji et al. [128].

Partition	$Pearson{-}\rho$
0: Kirchberg	0.24
1: Cessange	0.63
2: Gare	0.53
3: Merl	0.58

Table 6.1: Real Data Study: Pearson correlation between estimated and real traffic states by partition

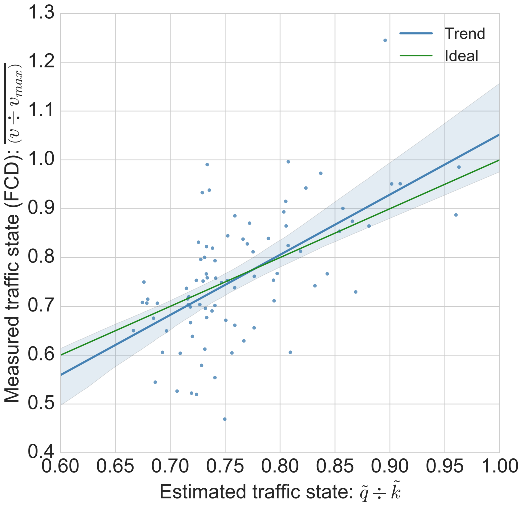


Figure 6.9: Real Data Study: Mobile data-based traffic state predictions vs ground-truth floating car data

In Fig. 6.9, we can see the scatter of predicted and true traffic states. We can see that the error appears to be independent from the traffic state, which is likely due to noise in the FCD and the relatively low resolution of our data set. The *Mean Absolute Percentage Error* (MAPE) amounts to 10.02%, which is comparable to other mobile network-based traffic state estimation techniques [6], but in an urban setting and with a much more interpretable model.

Limitations

The main limitation of our data sets is the temporal aggregation resolution of 1 hour of the mobile network data set. However, we are confident that the results will transfer onto higher temporal resolutions, and will go into some of the specific reasons of this in the following section. Another limitation is the low amount of congestion observed in both studies. However, in previous work, we observed that in deliberately congested situations, there is evidence that the mobile network data also reflects low-throughput/high-density situations [Derrmann et al., 2017a].

The results achieved in this work considering urban areas are much better than previous work on real-world data [Derrmann et al., 2017c], where we encountered very low correlation in some of the non-highway partitions. Thus, we believe that the polynomial density model introduced in this work is the key to adequately estimating urban MFDs from mobile network signaling data.

6.2.5 Discussion

The main promising result of this work is that even in complex, realistic networks with heterogeneous zones and unequally spaced mobile base stations, MFDs emerge from mobile phone data. The density proxy functions and MFDs we computed proved to show significant predictive power, leading to a MAPE of 11.12% on real data, which can compete with prior studies on (less complex) highway scenarios [6, 90, 125].

As expected, the real data prediction errors exceed those from the simulation run. This is due to the various limitations and simplified aspects in the simulation, avoiding e.g. stationary users and ping-pong handovers. Generally speaking, the simulated results were surprisingly similar in prediction quality to the real-data ones, which gives rise to a promising direction for future work.

The most important questions that arise from this work are whether complete MFDs can be extracted from mobile network data if there is a significant amount of congestion, and which is the spatio-temporal scale this is possible.

Regarding the first of these questions, we have confirmed the emergence of flow-density proxy relationships similar to MFD in the uncongested and saturated phases, as partially observed in the simulation study and more clearly in the real data study. The fact that our urban study regions do not exhibit heavy congestion and thus do not produce the descending phase of the flow-density diagram is a limitation of this work. Higher density traffic conditions will have to be studied in future work, to allow comparing mobile data-based results with studies on loop detectors by Buisson et al. [24] and Geroliminis et al. [130]. It is critical to investigate how precisely density can be approximated with mobile network in low-throughput traffic conditions, to verify whether distinguishing between low and high density situations is possible. However, in a previous simulation study involving artificially high traffic demand, we have shown that there is some evidence that this is likely the case [Derrmann et al., 2017a], but it has yet to be shown using pure handover data.

Regarding the second question, Geroliminis and Daganzo [22] have indicated 10 km^2 as lower bound on the spatial scale for the emergence of MFD from conventional loop detector signals. Therefore, in this study, we opted for creating 4 partitions with an average area of 12km^2. However, it would be necessary to investigate the impact of partition size on the flow-density approximation and their variance, and to evaluate the temporal scale at which the traffic states can be reasonably estimated from handover counts when using real-world data.

Fig. 6.10 shows the differences between the MFDs generated from both studies. For this purpose, we scaled all flow and density proxies into the ranges $[0, 1]$ so as to be able to compare both curves. The difference between both studies' results is surprisingly low, as they follow similar, mostly linear trends in partitions 0 and 3, and approaching saturated states in partitions 1 and 2. This indicates that the impact of ping-pong handovers, pedestrians and stationary users is not as high as feared, supporting the utility of handover data for mobility studies. Both studies also yielded comparable space-mean density values $\overline{\rho} \in [0.015, 0.05]\ \frac{veh.}{m}$. These mean density values also indicate that the observed congestion is not severe or covering the majority of any partition, as can also be seen from the plots in Fig. 6.10 that reach the saturated but not the descending phase.

While there are some differences between reality and simulation, we could show the predictive power of MFDs in both studies. However, we could also observe the need to partition the network in a more homogeneity-based approach. While in past studies [**?**, **?**], we focused on clustering the mobile network based on handovers and partitioning the road network according to these clusters, we have now identified that going in the opposite direction (road network first) is a more promising method. Different approaches of partitioning road networks into homogeneous partitions have been published and are being used [128,131,132]. The methodology in this study would certainly benefit from such improved partitioning, and we are convinced that with higher temporal resolution data and homogeneity-focused spatial partitioning, the correlation between road and mobile networks will be even stronger.

We are convinced that with higher temporal resolution data and spatial partitioning conditioned on homogeneity, the correlation between road and mobile networks will even be stronger. The potential impact for ITS applications of enabling mobile networks as a data source in this way is vast. In particular, our results pave the way for future studies that might generalize our findings to other cities and confirm mobile network signaling data as an invaluable source of insights for transportation engineers, either as a complement to existing solutions or even as the sole predictor. Urban areas in developing countries could

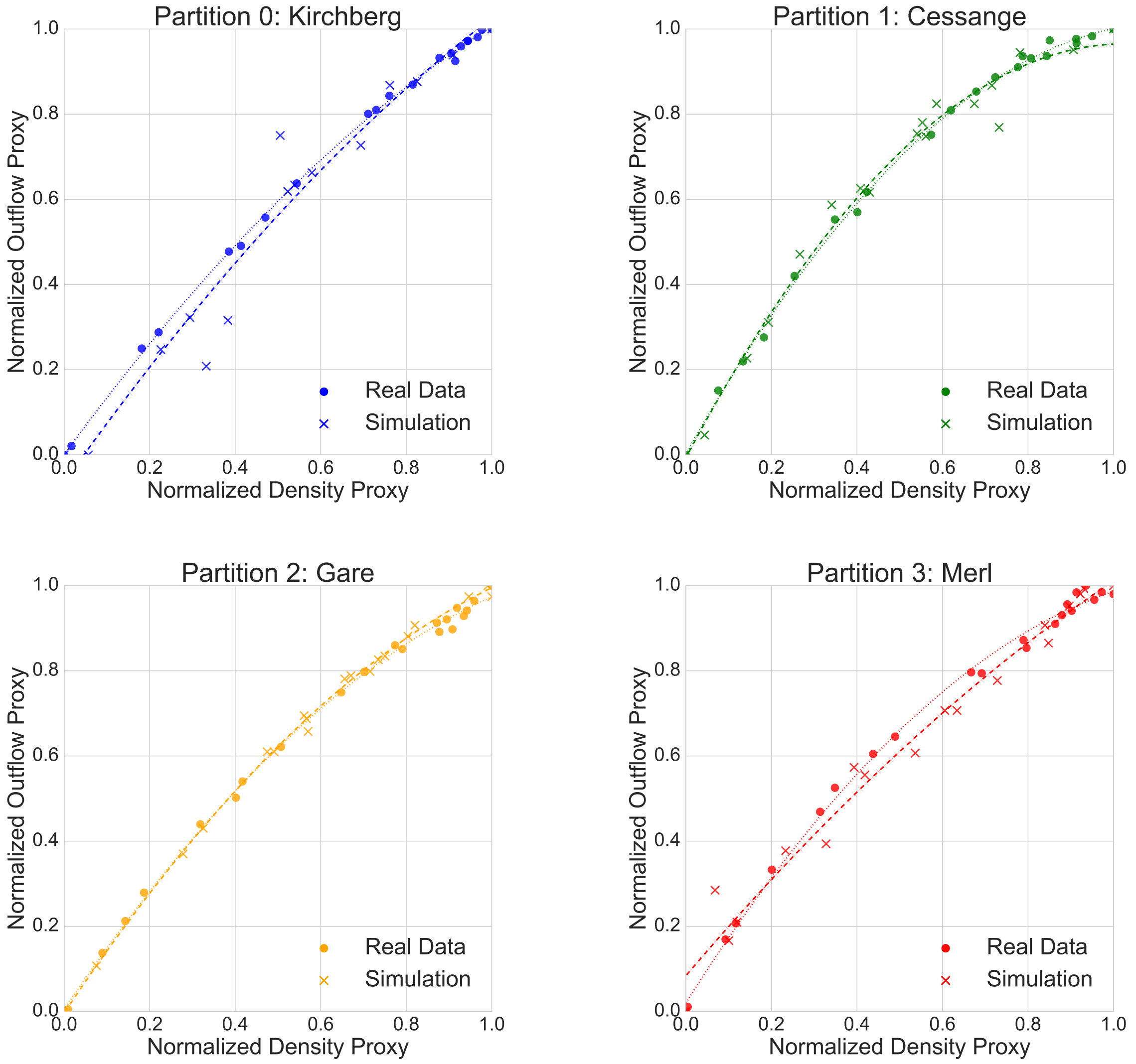


Figure 6.10: Comparison of Mobile Network MFD approximations: Normalized Flow-density relationships by partition

benefit particularly from this, in cases where mobile network infrastructure is relatively good and it is desirable to introduce or improve traffic sensing.

Chapter 7

Conclusions

In this dissertation, we have explored the links between mobile and transportation networks. We have used different kinds of mobile network data, and shown that they can indeed be used to model transportation supply and demand factors. We focused on data from the entire mobile network, which was collected using standard equipment. Specifically, we used both Call Detail Record (CDR) data and signalling data, more specifically handover statistics. Such data is readily available to the Mobile Network Operator (MNO), and although it contains information on stationary users as well as the moving population, we have demonstrated its utility in this dissertation. Both transportation supply and demand models proved to benefit from both types of mobile network data.

Before using real mobile network data, we created a simulation scenario that runs on VeinsLTE [41,68], linking the road and mobile networks of Luxembourg City. This gave us first insight into the possibilities that mobile network data offers, in particular signalling data in the form of aggregate handovers. Based on these findings, we continued with real data studies on data from Luxembourg. In the following section, we present the main conclusions of our simulation and real-data studies with respect to our research question, and subsequently discuss the conclusions from the different models we studied, along with relevant privacy considerations. Finally, we give perspectives for future work.

7.1 General Conclusions

Initially, we set out to answer the following research question:

How do mobile and transportation network behaviours correlate, and how can we leverage their interplay for transportation applications?

The following are the main conclusions we have drawn, with respect to this question, from the studies in this dissertation:

- *The link between the behaviour of mobile and road networks is strong, and it can be leveraged to model macroscopic traffic phenomena from mobile network data.* In all of our studies, we were able to identify strong links between mobile network data and the underlying population mobility. In Chapters 3 and 4, we showed that the square of the mean Cell Dwell Time (CDT) and travel time are proportional both in the case of Luxembourg City and of our synthetic data. This effect was present at a large spatial aggregation scale. At a smaller scale, a similar link between mobile and road networks was explored in Chapter 6. We showed that handovers in mobile networks are strong predictors for flows and density in partitions of the network. Based on these findings, we transferred the concept of Macroscopic Fundamental Diagrams (MFDs) into the domain of mobile network data.
- *Simulating vehicles only is sufficient for modelling macroscopic phenomena and the interaction between mobile and road networks. The bias brought about by low-mobility and stationary users is negligible when looking at large-scale traffic characteristics.* In Chapter 6, we were able to show that the Macroscopic Fundamental Diagrams (MFDs) generated from simulation and real data were almost identical. This is in line with the motivation of the work, i.e. to reduce the impact of low-mobility users (e.g. pedestrians) by considering handovers spanning large partitions of the mobile and road networks. Handover data was also used in Chapter 5, enabling the generation of an informative seed matrix of the flows entering and exiting Luxembourg City. Thus, in demand modelling, the effect of low-mobility users is also negligible. With respect to the research question, this means primarily that it is possible to simulate road traffic ITS solutions that rely on mobile network data, without the direct need to model all the modes. Our assumption that low-mobility users produce fewer handovers holds, as our methods are based on spatio-temporal aggregation and model only macroscopic traffic phenomena. Thus, low-mobility users and ping-pong handovers only contribute a small noise factor to our observed traffic variables.

- *Privacy can be preserved through the use of adequate aggregation schemes.* We were able to show – in all of the studies described in this dissertation – that aggregated data is sufficient to derive traffic parameters and fit models to the macroscopic behaviour of traffic. This is a significant result, as we have shown that no information on individual users needs to be shared by the MNO in order to publish models of the large-scale phenomena that transportation professionals are interested in for many of their applications. Instead, they only need to share the model parameters learned from their data (e.g. CDT distribution parameters or aggregated handover statistics). In related work, it was shown that for some technologies such as Bluetooth and Wireless LAN, unique identifiers expose a lot of information [133,134]. Few recurring observations of an individual – in some cases as few as four – can be sufficient to uniquely recognize that individual [63,64]. From a privacy perspective, the way that Bluetooth and WLAN work is potentially more concerning than mobile networks when it comes to passive data collection by individuals. This may change in the future, however, with direct connectivity between LTE-D2D enabled mobile devices. Generally speaking, when it comes to user trajectory data, it is essential to be wary of the potential implications [66]. Using the aggregation methods and models presented in this dissertation, these problems can be avoided, as MNOs need only share summarized parameters and aggregate statistics.

In the following sections, represented in chronological order as featured in this thesis, we discuss the specific conclusions of the studies concerning travel time, demand and traffic state modelling. We also present some conclusions on privacy and give directions for future related work.

7.2 Travel Time Modeling

In Chapter 4, we showed the potential of using CDT distributions for urban travel time prediction. We proposed a simple model of dwell times by cell pair, thus constructing a graph of dwell time distributions conditioned on the source cell. These findings build upon those by Janecek et al. [90] for highways. However, as we were limited to using synthetic data, the results are preliminary: they are to be confirmed with real data, where the distributions must be sliced so as to separate vehicular and stationary users. We also produced visualizations for the dwell time graphs, allowing quick assessment of which cell transitions are slowed down relative to their usual state. In theory, dwell time based travel time estimation can work well, but in practice, obtaining the necessary data can be costly, as this type of data is not as readily available as handover counts or CDRs. Consequently we decided to focus on traffic states based on handover counts in our later studies.

At a large aggregation scale, we found proportionality between the squared average of dwell times and the number of handovers in our synthetic data set, that was present in both the studies in Chapters 3 and 4. This further motivated us to look into handover data in order to model traffic states, and led to the research in Chapter 6.

7.3 Demand Modelling

In Chapter 5, we proposed a methodology for imputing intermediate user positions into Call Detail Records (CDR) data, and subsequently fitting a model of cell dwell times to the imputed data. The resulting cell dwell times lend themselves for the temporal component in the trip generation and distribution phases of a four-step demand estimation model. Our methodology, which we applied to the D4D Senegal data set, yielded data which was very similar in distribution to real-world cell dwell time data (up to a 5% shift in cumulative probability across the full data range). While the exact corresponding validation data was missing, we were able to show that the model-produced statistics correspond to results from other studies on different countries. In this context, future work will need to show whether or not the resulting dwell times from CDR-based models are realistic enough for demand estimation, but the results of this study strongly suggest that this is the case.

In another study, we demonstrated that aggregated handovers can serve as a valuable, privacy-friendly tool when estimating dynamic traffic flows between zones. Mobile phone data can help with initializing valid seed Origin-Destination (O-D) matrices, which allow a faster convergence of demand estimation algorithms. In this study, we observed a significant speed-up of the estimation of the O-D matrices, supporting the idea that reasonable initial values for the seed matrix can be derived from mobile phone data.

7.4 Traffic State Modelling

We have shown that there is significant similarity between simulated and actual traffic state models in Chapter 6, supporting the claim that the LuST-LTE scenario [Derrmann et al., 2016a] is a useful tool for simulating the interplay between the road and mobile phone networks. Following the simulation study, we proceeded to using real handover data for the traffic state estimation problem. To do this, we proposed a novel methodology to link mobile network signalling data to the underlying road network. We showed that it is possible to compute approximations of the road network partitions' *Macroscopic Fundamental Diagrams* (MFDs) using only aggregated mobile phone handover counts. To the best of our knowledge, ours is the first work to show that this link exists and that it can be reliable for real-world data in urban areas.

We first evaluated our methodology in a simulation study, which was limited by the absence of pedestrians and the difficulty of adequately simulating the demand in the network, yielding a low Mean Absolute Percentage Error (MAPE) of 10.2% in prediction. We then generalized our findings using the corresponding real-world data sets. While estimating the traffic states observed in the *Floating Car Data* (FCD) using mobile signaling data from an LTE network covering Luxembourg City, we achieved a MAPE of 11.12%, which compares well with previous studies (even those focusing on highways only) [6,90,125], but with the added advantage of being a simple, easily-interpretable model. The interpretability stems from the fact that the model only uses partitions' internal and exiting handovers as aggregate measures, and yields approximate measures for space-mean density, velocity and flows. The approximated MFDs exhibit low variance with respect to a concave flow-density function, which is in line with previous theoretical results on MFDs [130].

We also compared the resulting flow-density relationships of the simulation and real-data studies, and were able to show that they match and that the effect of the absence of pedestrians and stationary users from the simulation was negligible. These results are very encouraging as they show that the presented methodology is able to capture the traffic dynamics independently from the moving-to-stationary user ratio, at least in the low-to-moderate congestion situations given in Luxembourg City. The fact that pedestrian signals do not significantly influence flow-density approximation also lends more credibility to previous results of simulation studies, where we showed that mobile networks can also detect situations of low throughput and high traffic density [Derrmann et al., 2017a].

Using only two input variables from a set of base stations of a mobile network, it is possible to express their coverage area's traffic profile and make reasonably precise predictions of its traffic state. In this context, it is noteworthy that the predictive power that was achieved in this work is not its only quality.

One particular strength of this model is its privacy-friendliness, as it uses only aggregated data instead of data on individuals, thus avoiding a common pitfall of using mobility-related data, and mobile network data in particular.

Most importantly, this study has shown that the uncongested and saturation density regions of MFDs can be approximated using signalling data. Thus, it is desirable to find out whether highly congested networks exhibiting grid-lock phenomena could be approximated similarly. Such highly congested networks cover the full traffic density range of a classical MFD as shown e.g. by Geroliminis and Daganzo on the Yokohama network [22]. Note, however, that fully congested cases (i.e. gridlock) is not observed in most urban road networks, and that our results for Luxembourg City resemble those of other cities, e.g, Toulouse [24] and Brisbane [135]. Urban freeways also do not always produce strongly congested phase MFDs, e.g. Minneapolis, Chicago and Portland, as described by Saberi et al. in [136].

Currently, transportation researchers are actively looking for novel ways of obtaining MFDS, because they enable various planning and control measures. In this vein, there are active initiatives looking for novel data sources that show the emergence of MFDs, e.g. the MFD Dataquest [137]. We believe that our work is a first indicator that mobile network signaling data is a potential candidate to be such a data source, and this line of research should be continued for other networks to confirm our findings at a higher temporal resolution.

7.5 Privacy Considerations

The key element to leveraging communication network data in a privacy preserving manner is to perform adequate aggregation that suits the requirements of the target application. Ideally, this is performed close to the source, avoiding the exposure of sensitive data beyond the data provider's facilities. For some applications, fully aggregating data can be sufficient, as we have shown e.g. for the MFD models computed from aggregated mobile handover counts. In terms of privacy, the most important concept that we followed in the studies in this dissertation is that a MNO only shares or publishes aggregated data and the parameters of statistical models. E.g., for the CDT model, these are the parameters of the dynamic dwell time distributions, and for the MFD models these are the density model coefficients and the aggregate number of inter-cell handovers. This means no direct information regarding an individual mobile network subscriber needs to be shared outside the MNO. Instead, the data is aggregated over a large area or number of subscribers.

In the studies presented in this dissertation, we have worked on data that was aggregated in both the temporal and spatial domains. In the spatial dimension, it is important to know which model will be used and at which scale this model performs best. For example, for the MFD model, a spatial scale of 10 km^2 per zone is reasonable according to the related literature [22]. Thus, the spatial aggregation step can greatly reduce the amount of data and thus the privacy footprint. Concerning the exact aggregation technique, in Chapter 3, we have confirmed that the Voronoi tesselation is a useful approximation of LTE cell coverage even if buildings and a realistic propagation model are taken into account. Thus, the wide adoption of this computationally-efficient method in mobile network data analysis is justified. The fact that this is a readily-available algorithm on most platforms allows a rapid deployment of spatial aggregation in this manner.

For mobility purposes, temporal aggregation is a crucial step. A high temporal resolution is very important for numerous applications, most importantly real-time traffic forecasting. While we have shown in Chapter 6 that hourly data can provide some value there is also significant interest in having a higher temporal resolution. In other settings, the temporal component of the data can be rounded to the closest anchor point; for example, the D4D challenge CDR data sets are truncated to 10-minute precision. In Chapter 5, based on one of the D4D data sets, we impute user trajectories and then proceed to aggregate that data across the entire user base, thus avoiding the privacy-critical characteristic of CDR data [63, 64, 66].

Essentially, no direct (meta-)data should leave a communication network operator's facilities; instead, following the aggregation process, the model-fitting process can be run at the data acquisition source. This ensures that the output data is simply the parameterization of a model, such as the dwell time distribution parameters in Chapter 5. A potential attacker has little use for mobility model parameters, while researchers and traffic engineers can use them for estimation or simulation purposes. It is important, however, to make sure that the aggregation zones (i.e. the spatial aggregation scale) are sufficiently large and densely populated to avoid re-identification [63, 66].

To summarize, we have shown that there are efficient aggregation methods that can enable mobile network data as a reliable and privacy-neutral data source for ITS applications. With suitable methods, as presented in this thesis, the data can be transformed within the perimeter of the providing network operator, and provide real value for mobile network operators and transportation agencies.

7.6 Future research perspectives

Using heterogeneous simulators like VeinsLTE [41] [68] and scenarios such as LuST-LTE, as presented in this thesis, there is a significant potential for running realistic studies on (potentially multimodal) mobility in a realistic setting in terms of demand and connectivity. One major application of this is traffic optimization, ranging from gating with traffic light control to incentive-based methods.

In general, the privacy and legal aspects of measuring mobility are important challenges, especially when considering the next generations of communication networks that will enable more precise user localization. There are significant challenges in anonymizing user data while still making use of the data [63, 66], and we believe that a legal framework within which MNOs can share their data would be beneficial to research and to society in general.

As for the results that we achieved with real data, future research can address the questions of spatiotemporal resolution. Eventually, we see great potential in a holistic traffic estimation approach combining exo- and endogenic communication data, i.e. the fusion of e.g. V2I and cellular data. This will allow an accurate picture of the distribution of stationary and moving users to be obtained. The traffic state model that we have proposed is very promising in the low-congestion network of Luxembourg City. Based on the work presented in this dissertation, new directions in the modelling of macroscopic traffic phenomena from mobile network data have emerged. We have shown its vast potential for smart city and transportation planning efforts with data available today, and we encourage mobile network operators to grasp the opportunity of making their data available for research efforts. We are convinced that for MNOs, sharing aggregated statistics and/or mobility models computed from their data with public agencies is a win-win situation, as both parties can profit from the increasing value and insights that today's mobile network data offers.

List of Tables

List of Figures

Bibliography

[1] International Telecommunication Union, “Measuring the information society,” 2017. [Online]. Available: https://www.itu.int/en/ITU-D/Statistics/Documents/publications/misr2016/MISR2016-w4.pdf

[2] “Cisco Visual Networking Index (VNI) White Paper: Global Mobile Data Traffic Forecast 2016–2021,” 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[3] United Nations Global Pulse, *Mobile Data for Development Primer*, 2013. [Online]. Available: http://www.unglobalpulse.org/sites/default/files/Mobile%20Data%20for%20Development%20Primer_Oct2013.pdf

[4] V. D. Blondel, M. Esch, C. Chan, F. Clérot, P. Deville, E. Huens, F. Morlot, Z. Smoreda, and C. Ziemlicki, “Data for Development: the D4D Challenge on Mobile Phone Data,” *CoRR*, vol. abs/1210.0137, 2012. [Online]. Available: http://arxiv.org/abs/1210.0137

[5] Telecom Italia, “TIM Big Data Challenge,” 2015. [Online]. Available: www.telecomitalia.com/bigdatachallenge

[6] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, “Large-scale mobile traffic analysis: a survey,” *Communications Surveys Tutorials, IEEE*, no. 99, 10 2015.

[7] D. Naboulsi, M. Fiore, and R. Stanica, “Human Mobility Flows in the City of Abidjan,” in *3rd International Conference on the Analysis of Mobile Phone Datasets*, Boston, United States, May 2013, pp. 1–8. [Online]. Available: https://hal.inria.fr/hal-00908277

[8] M. Berlingerio, F. Calabrese, G. Lorenzo, R. Nair, F. Pinelli, and M. L. Sbodio, *AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, vol. 8190, ch. 50, pp. 663–666. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-40994-3_50

[9] M. Zilske and K. Nagel, "Building a minimal traffic model from mobile phone data," MIT (Cambridge, MA), Tech. Rep., May 2013, http://perso.uclouvain.be/vincent.blondel/netmob/2013/NetMob2013-program-v1.pdf. Also VSP WP 13-03, see www.vsp.tu-berlin.de/publications.

[10] A. Culotta, "Towards detecting influenza epidemics by analyzing Twitter messages," in *Proceedings of the First Workshop on Social Media Analytics*, ser. SOMA '10. New York, NY, USA: ACM, 2010, pp. 115–122. [Online]. Available: http://doi.acm.org/10.1145/1964858.1964874

[11] F. Calabrese, L. Ferrari, and V. D. Blondel, "Urban sensing using mobile phone network data: a survey of research," *ACM Computing Surveys (csur)*, vol. 47, no. 2, p. 25, 2015.

[12] E. Uhlemann, "Continued dispute on preferred vehicle-to-vehicle technologies [connected vehicles]," *IEEE Vehicular Technology Magazine*, vol. 12, no. 3, pp. 17–20, 2017.

[13] S. A. D. Donna, G. Cantelmo, and F. Viti, "A markov chain dynamic model for trip generation and distribution based on CDR," in *Proceedings of the 2015 IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), Budapest, Hungary, June 3-5, 2015*, 2015, pp. 243–250. [Online]. Available: https://doi.org/10.1109/MTITS.2015.7223263

[14] Y. Sheffi, *Urban Transportation Networks: Equilibrium Analysis With Mathematical Programming Methods*, 01 1984.

[15] J. d. D. Ortzar and L. Willumsen, *Modelling Transport, Fourth Edition*, 06 2011.

[16] E. Cascetta, *Transportation systems analysis. Models and applications. 2nd ed*, 01 2009, vol. 29.

[17] M. G. McNally, “The four-step model,” in *Handbook of Transport Modelling: 2nd Edition*. Emerald Group Publishing Limited, 2007, pp. 35–53.

[18] T. Toledo, T. Kolechkina, P. Wagner, B. Ciuffo, C. Azevedo, V. Marzano, and G. Flötteröd, “Network model calibration studies,” *Traffic Simulation and Data: Validation Methods and Applications*, p. 141, 2014.

[19] C. Antoniou, J. Barcel, M. Breen, M. Bullejos, J. Casas, E. Cipriani, B. Ciuffo, T. Djukic, S. Hoogendoorn, V. Marzano, L. Montero, M. Nigro, J. Perarnau, V. Punzo, T. Toledo, and H. van Lint, “Towards a generic benchmarking platform for origindestination flows estimation/updating algorithms: Design, demonstration and validation,” *Transportation Research Part C: Emerging Technologies*, vol. 66, no. Supplement C, pp. 79 – 98, 2016, advanced Network Traffic Management: From dynamic state estimation to traffic control. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X15003101

[20] D. B. Work, O.-P. Tossavainen, S. Blandin, A. M. Bayen, T. Iwuchukwu, and K. Tracton, “An ensemble kalman filtering approach to highway traffic estimation using gps enabled mobile devices,” in *Proceedings of the 47th IEEE Conference on Decision and Control (CDC) 2008*. IEEE, 2008, pp. 5062–5068.

[21] B. Greenshields, W. Channing, H. Miller *et al.*, “A study of traffic capacity,” in *Highway research board proceedings*, vol. 1935. National Research Council (USA), Highway Research Board, 1935.

[22] N. Geroliminis and C. F. Daganzo, “Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings,” *Transportation Research Part B: Methodological*, vol. 42, no. 9, pp. 759–770, 2008.

[23] H. S. Mahmassani, M. Saberi, and A. Zockaie, “Urban network gridlock: Theory, characteristics, and dynamics,” *Transportation Research Part C: Emerging Technologies*, vol. 36, pp. 480–497, 2013.

[24] C. Buisson and C. Ladier, “Exploring the impact of homogeneity of traffic measurements on the existence of macroscopic fundamental diagrams,” *Transportation Research Record: Journal of the Transportation Research Board*, no. 2124, pp. 127–136, 2009.

[25] N. Geroliminis, J. Haddad, and M. Ramezani, "Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: A model predictive approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 348–359, 2013.

[26] J. Haddad, M. Ramezani, and N. Geroliminis, "Cooperative traffic control of a mixed network with two urban regions and a freeway," *Transportation Research Part B: Methodological*, vol. 54, pp. 17–36, 2013.

[27] R.-P. Schäfer, K.-U. Thiessenhusen, E. Brockfeld, and P. Wagner, "A traffic information system by means of real-time floating-car data," 2002.

[28] B. M. Williams and L. A. Hoel, "Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results," *Journal of transportation engineering*, vol. 129, no. 6, pp. 664–672, 2003.

[29] J.-S. Yang, "Travel time prediction using the gps test vehicle and kalman filtering techniques," in *American Control Conference, 2005. Proceedings of the 2005.* IEEE, 2005, pp. 2128–2133.

[30] Y. Liu, Y. Wang, X. Yang, and L. Zhang, "Short-term Travel Time Prediction by Deep Learning: A Comparison of Different LSTM-DNN Models," in *Proceedings of the 2017 IEEE Intelligent Transportation Systems Conference (ITSC).* IEEE, 2017, pp. 2083–2090.

[31] Y. Hou, P. Edara, and Y. Chang, "Road Network State Estimation Using Random Forest Ensemble Learning," in *Proceedings of the 2017 IEEE Intelligent Transportation Systems Conference (ITSC).* IEEE, 2016, pp. 2091–2096.

[32] A. Hofleitner, R. Herring, P. Abbeel, and A. Bayen, "Learning the dynamics of arterial traffic from probe data using a dynamic bayesian network," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1679–1693, 2012.

[33] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of SUMO - Simulation of Urban MObility," *International Journal On Advances in Systems and Measurements*, vol. 5, no. 3&4, pp. 128–138, December 2012.

[34] K. S. Lee, J. K. Eom, and D.-s. Moon, “Applications of transims in transportation: A literature review,” *Procedia Computer Science*, vol. 32, pp. 769–773, 2014.

[35] A. Horni, K. Nagel, and K. W. Axhausen, *The multi-agent transport simulation MATSim.*

[36] I. Kaddoura, B. Kickhöfer, A. Neumann, and A. Tirachini, “Optimal public transport pricing: Towards an agent-based marginal social cost approach,” *Journal of Transport Economics and Policy (JTEP)*, vol. 49, no. 2, pp. 200–218, 2015.

[37] T. Novosel, L. Perković, M. Ban, H. Keko, T. Pukšec, G. Krajačić, and N. Duić, “Agent based modelling and energy planning–utilization of matsim for transport energy demand modelling,” *Energy*, vol. 92, pp. 466–475, 2015.

[38] B. Yao, P. Hu, X. Lu, J. Gao, and M. Zhang, “Transit network design based on travel time reliability,” *Transportation Research Part C: Emerging Technologies*, vol. 43, no. Part 3, pp. 233 – 248, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X13002647

[39] S. Galland, L. Knapen, N. Gaud, D. Janssens, O. Lamotte, A. Koukam, G. Wets *et al.*, “Multi-agent simulation of individual mobility behavior in carpooling,” *Transportation Research Part C: Emerging Technologies*, vol. 45, pp. 83–98, 2014.

[40] W. He, K. Hwang, and D. Li, “Intelligent carpool routing for urban ridesharing by mining gps trajectories,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2286–2296, 2014.

[41] C. Sommer, R. German, and F. Dressler, “Bidirectionally Coupled Network and Road Traffic Simulation for Improved IVC Analysis,” *IEEE Transactions on Mobile Computing*, vol. 10, no. 1, pp. 3–15, January 2011.

[42] B. Sliwa, J. Pillmann, F. Eckermann, and C. Wietfeld, “Limosim: A lightweight and integrated approach for simulating vehicular mobility with omnet++,” in *OMNeT++ Community Summit 2017*, Bremen, Germany, Sep 2017.

[43] I. Turcanu, C. Sommer, A. Baiocchi, and F. Dressler, “Pick the right guy: Cqi-based lte forwarder selection in vanets,” in *2016 IEEE Vehicular Networking Conference (VNC)*, Dec 2016, pp. 1–8.

[44] L. Codeca, R. Frank, and T. Engel, "Improving traffic in urban environments applying the wardrop equilibrium," in *Network Protocols (ICNP), 2013 21st IEEE International Conference on.* IEEE, 2013, pp. 1–6.

[45] M. Hadiuzzaman and T. Z. Qiu, "Cell transmission model based variable speed limit control for freeways," *Canadian Journal of Civil Engineering*, vol. 40, no. 1, pp. 46–56, 2013.

[46] M. Forster, R. Frank, M. Gerla, and T. Engel, "A cooperative advanced driver assistance system to mitigate vehicular traffic shock waves," in *Proceedings of the 2014 IEEE INFOCOM.* IEEE, 2014, pp. 1968–1976.

[47] M. W. Levin and S. D. Boyles, "A cell transmission model for dynamic lane reversal with autonomous vehicles," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 126–143, 2016.

[48] C. E. Cortés, D. Sáez, F. Milla, A. Núñez, and M. Riquelme, "Hybrid predictive control for real-time optimization of public transport systems operations based on evolutionary multi-objective optimization," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 5, pp. 757–769, 2010.

[49] M. Dessouky, R. Hall, L. Zhang, and A. Singh, "Real-time control of buses for schedule coordination at a terminal," *Transportation Research Part A: Policy and Practice*, vol. 37, no. 2, pp. 145 – 164, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0965856402000101

[50] D. Beymer, P. McLauchlan, B. Coifman, and J. Malik, "A real-time computer vision system for measuring traffic parameters," in *Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* IEEE, 1997, pp. 495–501.

[51] J. Barceló, L. Montero, L. Marqués, and C. Carmona, "Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2175, pp. 19–27, 2010.

[52] C. Bachmann, M. J. Roorda, B. Abdulhai, and B. Moshiri, "Fusing a bluetooth traffic monitoring system with loop detector data for improved freeway traffic speed

estimation," *Journal of Intelligent Transportation Systems*, vol. 17, no. 2, pp. 152–164, 2013.

[53] M. R. Friesen and R. D. McLeod, "Bluetooth in intelligent transportation systems: A survey," *International Journal of Intelligent Transportation Systems Research*, vol. 13, no. 3, pp. 143–153, Sep 2015. [Online]. Available: https://doi.org/10.1007/s13177-014-0092-1

[54] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, 2011.

[55] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport origindestination matrix from passive smartcard data from santiago, chile," *Transportation Research Part C: Emerging Technologies*, vol. 24, no. Supplement C, pp. 9 – 18, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X12000095

[56] T. Kusakabe and Y. Asakura, "Behavioural data mining of transit smart card data: A data fusion approach," *Transportation Research Part C: Emerging Technologies*, vol. 46, no. Supplement C, pp. 179 – 191, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X14001612

[57] X. Ma, Y.-J. Wu, Y. Wang, F. Chen, and J. Liu, "Mining smart card data for transit riders travel patterns," *Transportation Research Part C: Emerging Technologies*, vol. 36, no. Supplement C, pp. 1 – 12, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X13001630

[58] V. Kumar, L. Lin, D. Krajzewicz, F. Hrizi, O. Martinez, J. Gozalvez, and R. Bauza, "itetris: Adaptation of its technologies for large scale integrated simulation," in *Proceedings of the 71st IEEE Vehicular Technology Conference (VTC 2010-Spring)*. IEEE, 2010, pp. 1–5.

[59] W. Bronzi, T. Derrmann, G. Castignani, and T. Engel, "Towards characterizing bluetooth discovery in a vehicular context," in *Proceedings of the 2016 IEEE Vehicular Networking Conference (VNC)*, Dec 2016, pp. 1–4.

[60] N. Shlayan, A. Kurkcu, and K. Ozbay, *Exploring pedestrian bluetooth and WiFi detection at public transportation terminals.* United States: Institute of Electrical and Electronics Engineers Inc., 12 2016, pp. 229–234.

[61] P. Lescuyer and T. Lucidarme, *Evolved Packet System (EPS): The LTE and SAE Evolution of 3G UMTS.* Wiley Publishing, 2008.

[62] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in rome," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 1, pp. 141–151, 2011.

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

[64] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. ". Pentland, "Unique in the shopping mall: On the reidentifiability of credit card metadata," *Science*, vol. 347, no. 6221, pp. 536–539, 2015. [Online]. Available: http://science.sciencemag.org/content/347/6221/536

[65] J. Petit, F. Schaub, M. Feiri, and F. Kargl, "Pseudonym schemes in vehicular networks: A survey," *IEEE communications surveys & tutorials*, vol. 17, no. 1, pp. 228–255, 2015.

[66] M. Gramaglia, M. Fiore, A. Tarable, and A. Banchs, "Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories," in *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, May 2017, pp. 1–9.

[67] L. Codecà, R. Frank, S. Faye, and T. Engel, "Luxembourg SUMO Traffic (LuST) Scenario: Traffic Demand Evaluation," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 2, pp. 52–63, 2017.

[68] F. Hagenauer, F. Dressler, and C. Sommer, "Poster: A simulator for heterogeneous vehicular networks," in *Proceedings of the 2014 IEEE Vehicular Networking Conference (VNC)*, Dec 2014, pp. 185–186.

[69] S. Uppoor, O. Trullols-Cruces, M. Fiore, and J. M. Barcelo-Ordinas, "Generation and analysis of a large-scale urban vehicular mobility dataset," *IEEE Transactions on Mobile Computing*, vol. 13, no. 5, pp. 1061–1075, May 2014.

[70] M. Rondinone, J. Maneros, D. Krajzewicz, R. Bauza, P. Cataldi, F. Hrizi, J. Gozalvez, V. Kumar, M. Röckl, L. Lin *et al.*, "itetris: a modular simulation platform for the large scale evaluation of cooperative its applications," *Simulation Modelling Practice and Theory*, vol. 34, pp. 99–125, 2013.

[71] A. Varga *et al.*, "The OMNeT++ discrete event simulation system," in *Proceedings of the European simulation multiconference (ESM2001)*, vol. 9, no. S 185. sn, 2001, p. 65.

[72] A. Virdis, G. Stea, and G. Nardini, *Simulating LTE/LTE-Advanced Networks with SimuLTE.* Springer International Publishing, 2015, pp. 83–105.

[73] S. Faye and C. Chaudet, "Characterizing the Topology of an Urban Wireless Sensor Network for Road Traffic Management," *IEEE Transactions on Vehicular Technology*, 2015.

[74] ——, "Connectivity analysis of wireless sensor networks deployments in smart cities," in *The 22nd IEEE Symposium on Communications and Vehicular Technology in the Benelux (SCVT)*, Luxembourg City, Luxembourg, Nov. 2015.

[75] D. Puccinelli, O. Gnawali, S. Yoon, S. Santini, U. Colesanti, S. Giordano, and L. Guibas, "The impact of network topology on collection performance," in *8th European conference on Wireless sensor networks*, 2011.

[76] T. Ducrocq, M. Hauspie, N. Mitton, and S. Pizzi, "On the Impact of Network Topology on Wireless Sensor Networks Performances Illustration with Geographic Routing," in *International Workshop on the Performance Analysis and Enhancement of Wireless Networks (PAEWN)*, Victoria, Canada, May 2014.

[77] L. Codecà, R. Frank, and T. Engel, "Luxembourg sumo traffic (lust) scenario: 24 hours of mobility for vehicular networking research," in *Proceedings of the 2015 IEEE Vehicular Networking Conference (VNC)*, Dec 2015, pp. 1–8.

[78] S. Uppoor and M. Fiore, "Characterizing Pervasive Vehicular Access to the Cellular RAN Infrastructure: An Urban Case Study," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 6, pp. 2603–2614, June 2015.

[79] M. Series, "Guidelines for evaluation of radio interface technologies for IMT-Advanced."

[80] W. C. Jakes and D. C. Cox, *Microwave mobile communications.* Wiley-IEEE Press, 1994.

[81] F. Afroz, R. Subramanian, R. Heidary, K. Sandrasegaran, and S. Ahmed, "SINR, RSRP, RSSI and RSRQ Measurements in Long Term Evolution Networks," *International Journal of Wireless & Mobile Networks*, vol. 7, pp. 113–123, Aug 2015.

[82] R. Herring, A. Hofleitner, P. Abbeel, and A. Bayen, "Estimating arterial traffic conditions using sparse probe data," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*, Sept 2010, pp. 929–936.

[83] C. Nanthawichit, T. Nakatsuji, and H. Suzuki, "Application of probe-vehicle data for real-time traffic-state estimation and short-term travel-time prediction on a freeway," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1855, pp. 49–59, 2003.

[84] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Large-scale mobile traffic analysis: A survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 1, pp. 124–161, Firstquarter 2016.

[85] D. Valerio, A. D. Alconzo, F. Ricciato, and W. Wiedermann, "Exploiting cellular networks for road traffic estimation: a survey and a research roadmap," in *Vehicular Technology Conference, 2009. VTC Spring 2009. IEEE 69th.* IEEE, 2009, pp. 1–5.

[86] J. Reades, F. Calabrese, A. Sevtsuk, and C. Ratti, "Cellular census: Explorations in urban data collection," *Pervasive Computing, IEEE*, vol. 6, no. 3, pp. 30–38, 2007.

[87] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-time urban monitoring using cell phones: A case study in Rome," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 1, pp. 141–151, March 2011.

[88] R. Trasarti, A.-M. Olteanu-Raimond, M. Nanni, T. Couronn, B. Furletti, F. Giannotti, Z. Smoreda, and C. Ziemlicki, "Discovering urban and country dynamics from mobile phone data with spatial correlation patterns," *Telecommunications Policy*, vol. 39, no. 34, pp. 347 – 362, 2015, mobile phone data and geographic modelling. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0308596113002012

[89] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti, "Interactive visual clustering of large collections of trajectories," *VAST*, 2009.

[90] A. Janecek, D. Valerio, K. A. Hummel, F. Ricciato, and H. Hlavacs, "The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2551–2572, Oct 2015.

[91] J. Schlaich, T. Otterstatter, and M. Friedrich, "Generating trajectories from mobile phone data," in *Transportation Research Board 89th Annual Meeting*, no. 10-0374, 2010.

[92] S. Uppoor and M. Fiore, "Characterizing pervasive vehicular access to the cellular RAN infrastructure: An urban case study," *Vehicular Technology, IEEE Transactions on*, vol. 64, no. 6, pp. 2603–2614, June 2015.

[93] H. Bar-Gera, "Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel," *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 6, pp. 380 – 391, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0968090X07000393

[94] R. Chrobok, O. Kaumann, J. Wahle, and M. Schreckenberg, "Three categories of traffic data: Historical, current, and predictive," in *Proceedings of the 9th IFAC Symposium Control in Transportation Systems*, 2000, pp. 250–255.

[95] S. Holl and H. Plum, "PostGIS," *GeoInformatics*, vol. 03/2009, pp. 34–36, Apr. 2009. [Online]. Available: http://fluidbook.microdesign.nl/geoinformatics/03-2009/?page=34

[96] D. Kahle and H. Wickham, "ggmap: Spatial visualization with ggplot2," *The R Journal*, vol. 5, no. 1, pp. 144–161, 2013. [Online]. Available: http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

[97] P. V. Orlik and S. S. Rappaport, "A model for teletraffic performance and channel holding time characterization in wireless cellular communication with general session and dwell time distributions," *Selected Areas in Communications, IEEE Journal on*, vol. 16, no. 5, pp. 788–803, 1998.

[98] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, “Trajectory pattern mining,” in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2007, pp. 330–339.

[99] I. CIESIN, “Wri, 2000. gridded population of the world (gpw), version 2,” *Center for International Earth Science Information Network (CIESIN) Columbia University, International Food Policy Research Institute (IFPRI) and World Resources Institute (WRI), Palisades, NY.*

[100] S. A. Di Donna, G. Cantelmo, and F. Viti, “A Markov chain dynamic model for trip generation and distribution based on CDR,” in *Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2015 International Conference on.* IEEE, 2015, pp. 243–250.

[101] E. Halepovic and C. Williamson, “Characterizing and modeling user mobility in a cellular data network,” in *Proceedings of the 2Nd ACM International Workshop on Performance Evaluation of Wireless Ad Hoc, Sensor, and Ubiquitous Networks*, ser. PE-WASUN ’05. New York, NY, USA: ACM, 2005, pp. 71–78.

[102] S. Scepanovic, P. Hui, and A. Yla-Jaaski, “Revealing the Pulse of Human Dynamics in a Country from Mobile Phone Data,” in *D4D Challenge Submissions, NetMob*, 2013. [Online]. Available: http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf

[103] A. Hess, I. March, and D. Gillblad, “Exploring Communication and Mobility Behavior of 3G Network Users and Its Temporal Consistency,” in *Proceedings of the IEEE International Conference on Communications (ICC).* IEEE, 2015.

[104] K. S. Kung, K. Greco, S. Sobolevsky, and C. Ratti, “Exploring universal patterns in human home-work commuting from mobile phone data,” *PLoS ONE*, vol. 9, no. 6, p. e96180, 06 2014. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0096180

[105] M. Tizzoni, P. Bajardi, A. Decuyper, G. Kon Kam King, C. M. Schneider, V. Blondel, Z. Smoreda, M. C. Gonzez, and V. Colizza, “On the use of human mobility proxies for modeling epidemics,” *PLoS Comput Biol*, vol. 10, no. 7, p. e1003716, 07 2014. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pcbi.1003716

[106] D. Zhang, J. Huang, Y. Li, F. Zhang, C. Xu, and T. He, "Exploring human mobility with multi-source data at extremely large metropolitan scales," in *Proceedings of the 20th annual international conference on Mobile computing and networking.* ACM, 2014, pp. 201–212.

[107] Y.-B. Lin, S. Mohan, and A. Noerpel, "Queueing priority channel assignment strategies for pcs hand-off and initial access," *Vehicular Technology, IEEE Transactions on*, vol. 43, no. 3, pp. 704–712, Aug 1994.

[108] Y. Fang, "Hyper-erlang distribution model and its application in wireless mobile networks," *Wireless Networks*, vol. 7, no. 3, pp. 211–219, 2001. [Online]. Available: http://dx.doi.org/10.1023/A%3A1016617904269

[109] A. Corral-Ruiz, F. Cruz-Pérez, and G. Hernández-Valdez, *Cell Dwell Time and Channel Holding Time Relationship in Mobile Cellular Networks.* INTECH Open Access Publisher, 2012. [Online]. Available: http://books.google.lu/books?id=cs7hoAEACAAJ

[110] H. Hidaka, K. Saitoh, N. Shinagawa, and T. Kobayashi, "Statistical properties of measured vehicle motion and teletraffic in cellular communications," in *Multiaccess, Mobility and Teletraffic in Wireless Communications: Volume 5*, G. Stber and B. Jabbari, Eds. Springer US, 2000, pp. 291–303. [Online]. Available: http://dx.doi.org/10.1007/978-1-4757-5916-7_25

[111] H. Wang, F. Calabrese, G. Di Lorenzo, and C. Ratti, "Transportation mode inference from anonymized and aggregated mobile phone call detail records," in *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on.* IEEE, 2010, pp. 318–323.

[112] O. Jarv, R. Ahas, E. Saluveer, B. Derudder, and F. Witlox, "Mobile phones in a traffic flow: a geographical perspective to evening rush hour traffic analysis using call detail records," *PLoS One*, vol. 7, no. 11, 2012.

[113] M. R. Vieira, V. Frias-Martinez, N. Oliver, and E. Frias-Martinez, "Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics," in *Social Computing (SocialCom), 2010 IEEE Second International Conference on.* IEEE, 2010, pp. 241–248.

[114] A. Apolloni, A. Camacho, K. Eames, J. W. Edmunds, and S. Funk, "First steps for a Synthetic Population of Ivory Coast," in *D4D Challenge Submissions, NetMob*, 2013. [Online]. Available: http://perso.uclouvain.be/vincent.blondel/netmob/2013/D4D-book.pdf

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

[116] Autorité de Régulation des Télécommunications et des Postes, "Rapport Trimestriel sur le Marché des Télécommunications Juillet-Septembre 2014," http://www.artpsenegal.net/images/documents/Rapport%20T3_2014_Version_Finale.pdf, 2014.

[117] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang, "Human mobility patterns in cellular networks," *IEEE Communications Letters*, vol. 17, no. 10, pp. 1877–1880, October 2013.

[118] S. Uppoor and M. Fiore, "Characterizing pervasive vehicular access to the cellular RAN infrastructure: an urban case study," *Vehicular Technology, IEEE Transactions on*, vol. 64, no. 6, pp. 2603–2614, 2015.

[119] X. Liang, X. Zheng, W. Lv, T. Zhu, and K. Xu, "The scaling of human mobility by taxis is exponential," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 5, pp. 2135–2144, 2012.

[120] H. Hidaka, K. Saitoh, N. Shinagawa, and T. Kobayashi, "Terminal migration model in which cell dwell time is defined by different probability distributions in different cells," in *Multiaccess, Mobility and Teletraffic for Wireless Communications, volume 6*, X. Lagrange and B. Jabbari, Eds. Springer US, 2002, pp. 133–142.

[121] M. G. Demissie, G. Correia, and C. Bento, "Analysis of the pattern and intensity of urban activities through aggregate cellphone usage," *Transportmetrica A: Transport Science*, vol. 11, no. 6, pp. 502–524, 2015.

[122] G. Sagl, B. Resch, B. Hawelka, and E. Beinat, "From social sensor data to collective human behaviour patterns: Analysing and visualising spatio-temporal dynamics in urban environments," in *Proceedings of the GI-Forum*,

2012, pp. 54–63. [Online]. Available: http://senseable.mit.edu/papers/pdf/20121109_Sagl_etal_FromSocial_Forum2012.pdf

[123] K. Hui, C. Wang, and A. Kim, "Investigating the use of anonymous cellular phone data to determine intercity travel volumes and modes," *Transportation Research Board Annual Meeting*, 2017.

[124] H. Bar-Gera, "Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel," *Transportation Research Part C: Emerging Technologies*, vol. 15, no. 6, pp. 380–391, 2007.

[125] N. Caceres, L. M. Romero, F. G. Benitez, and J. M. del Castillo, "Traffic flow estimation models using cellular phone data," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1430–1441, 2012.

[126] V. V. Gayah, X. S. Gao, and A. S. Nagle, "On the impacts of locally adaptive signal control on urban network stability and the macroscopic fundamental diagram," *Transportation Research Part B: Methodological*, vol. 70, pp. 255–268, 2014.

[127] M. Saberi and H. Mahmassani, "Hysteresis and capacity drop phenomena in freeway networks: Empirical characterization and interpretation," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2391, pp. 44–55, 2013.

[128] Y. Ji and N. Geroliminis, "On the spatial partitioning of urban transportation networks," *Transportation Research Part B: Methodological*, vol. 46, no. 10, pp. 1639–1656, 2012.

[129] H. Dong, M. Wu, X. Ding, L. Chu, L. Jia, Y. Qin, and X. Zhou, "Traffic zone division based on big data from mobile phone base stations," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 278 – 291, 2015, big Data in Transportation and Traffic Engineering.

[130] N. Geroliminis and J. Sun, "Properties of a well-defined macroscopic fundamental diagram for urban traffic," *Transportation Research Part B: Methodological*, vol. 45, no. 3, pp. 605–617, 2011.

[131] C. Lopez, P. Krishnakumari, L. Leclercq, N. Chiabaut, and H. Van Lint, "Spatiotemporal partitioning of transportation network using travel time data," *Transportation*

Research Record: Journal of the Transportation Research Board, no. 2623, pp. 98–107, 2017.

[132] M. Saeedmanesh and N. Geroliminis, "Clustering of heterogeneous networks with directional flows based on snake similarities," *Transportation Research Part B: Methodological*, vol. 91, pp. 250–269, 2016.

[133] M. Vanhoef, C. Matte, M. Cunche, L. S. Cardoso, and F. Piessens, "Why mac address randomization is not enough: An analysis of wi-fi network discovery mechanisms," in *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security.* ACM, 2016, pp. 413–424.

[134] F.-L. Wong and F. Stajano, "Location privacy in bluetooth," *Security and privacy in ad-hoc and sensor networks*, pp. 176–188, 2005.

[135] T. Tsubota, A. Bhaskar, and E. Chung, "Macroscopic fundamental diagram for brisbane, australia: empirical findings on network partitioning and incident detection," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2421, pp. 12–21, 2014.

[136] M. Saberi and H. S. Mahmassani, "Empirical characterization and interpretation of hysteresis and capacity drop phenomena in freeway networks," *Transportation Research Record: Journal of the Transportation Research Board, Transportation Research Board of the National Academies, Washington, DC*, 2013.

[137] Subcommittee of AHB45 TRB Committee, "MFD Dataquest," 2016. [Online]. Available: https://sites.google.com/a/jltraffic.com/mfd-dataquest/home

Contributions

[Cantelmo et al., 2017] Cantelmo, G., Viti, F., and Derrmann, T. (2017). Effectiveness of the two-step dynamic demand estimation model on large networks. In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 356–361.

[Derrmann et al., 2016a] Derrmann, T., Faye, S., Frank, R., and Engel, T. (2016a). Poster: Lust-lte: A simulation package for pervasive vehicular connectivity. In *Vehicular Networking Conference (VNC), 2016 IEEE*, pages 1–2. IEEE.

[Derrmann et al., 2017a] Derrmann, T., Frank, R., Engel, T., and Viti, F. (2017a). How mobile phone handovers reflect urban mobility: A simulation study. In *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, pages 486–491.

[Derrmann et al., 2016b] Derrmann, T., Frank, R., Faye, S., Castignani, G., and Engel, T. (2016b). Towards privacy-neutral travel time estimation from mobile phone signalling data. In *2016 IEEE International Smart Cities Conference (ISC2)*, pages 1–6.

[Derrmann et al., 2017b] Derrmann, T., Frank, R., and Viti, F. (2017b). Towards estimating urban macroscopic fundamental diagrams from mobile phone signaling data: A simulation study. In *Transportation Research Board Annual Meeting 2017*.

[Derrmann et al., 2017c] Derrmann, T., Frank, R., Viti, F., and Engel, T. (2017c). Estimating urban road traffic states using mobile network signaling data (to appear). In *20th IEEE International Conference on Intelligent Transportation (ITSC)*.

[Faye et al., 2017] Faye, S., Cantelmo, G., Tahirou, I., Derrmann, T., Viti, F., and Engel, T. (2017). Demo: Mamba: A platform for personalised multimodal trip planning. In *Vehicular Networking Conference (VNC), 2017 IEEE*.