Towards Passive Tracking and Analyses of Human Mobility at Population Scale

Dissertation

zur

Erlangung der naturwissenschaftlichen Dokdorwürde

(Dr. sc. nat.)

Vorgelegt der

Mathematisch-naturwissenschaftlichen Faktultät

der

Universität Zürich

von

Oliver Burkhard

von

Schwarzhäusern (BE)

Promotionskommission

Prof. Dr. Robert Weibel (Vorsitz)

Prof. Dr. Ross Purves

Dr. Martin Tomko

Zürich, 2019

Mathematisch-naturwissenschaftliche Fakultät
der Universität Zürich

Dissertation
Towards Passive Tracking and Analyses of Human Mobility at Population Scale

Author:
Oliver Burkhard
Geographic Information Systems Unit
Department of Geography
University of Zurich Winterthurerstrasse 190
CH-8057 Zürich
Switzerland

http://www.geo.uzh.ch/en/units/gis.html

Summary

Understanding the movement of people can provide significant insights about both individuals and the societies they live in. If the *where*, *when*, *how* and *why* of movement are all known in detail, a wide variety of aspects about the people whose movement is observed can be revealed, ranging from their family structures, social lives, religious believes and consumer preferences. Given this wide range of potential insights, a wide variety of actors have shown interest in studying mobility, including academics, commercial entities and policy makers.

Perhaps most interested in the movement itself is the field of transportation research, that does not view movement just as a proxy for other aspects. Researchers in that domain have been asking question about the mobility behaviour of people for decades and have used many tools, from interviews to simulations, passenger counting and real life tracking to gain knowledge. Due to the importance of that knowledge some of those collection efforts are borne by the state, conducting expensive surveys to provide a good data basis from which many actors profit.

Great change, however has come in the availability of data. With the spread of smart phones that started a decade ago, staggering fractions of the population now carry a whole array of sensors – amongst them one for satellite tracking – with them wherever they go. With the more recent adaptation of the 4G standard, relatively precise localisation is now built into the system and this information is available to the service providers. In addition, with the advent of big data capabilities, large telephone providers can store this location based information that the smart phones of their customers reveal, resulting in large databases of information on human movement.

This thesis aims at contributing to understanding the implications of these developments in combination with advances in machine learning to facilitate the understanding of human movement at geographical scale, thus contributing to the aims of the various actors mentioned above. In particular, it proposes methods that are applicable to data from large fractions of the population and could thus help to increase the temporal and semantic resolution at which movements of entire populations observed.

First, a method is proposed to reconstruct the *when* and *where* of human movement based on very sparse – an average of about 3 pieces of information a day – passively sensed mobile phone data. The goal is to obtain estimations of the whereabouts of people during times where there is no signal to be observed. The method achieves this with the sole assumption of the repetitiveness of movement, constructing prototype days which are then used in the interpolation. The results shows that while the approximation is significantly better than using assumptions such as repetitive working weekdays for the interpolation, more detailed questions about the movement of the people itself, such as the mode of transport, may be difficult to answer from data that and spatially inaccurate.

The second part of this thesis tries to explore the extent of this problem by asking what accuracy and sampling rate are required from passive tracking systems to enable transport mode detection. In order to address this, GPS data are used, of which typically high spatial accuracy and sampling rate are assumed. These data are then subsampled and distorted to obtain data of worse and worse quality. With sampling rates and precisions that are comparable to what can be obtained today using passive tracking, transport mode detection is found to be feasible. Further improvements could be obtained by higher spatial accuracy, which is to be expected with the next generation of mobile phones.

The third part of this thesis explores a possible consequence of large scale transportation mode detection found possible by the second part. Knowing all the modes from passive tracking would allow for traffic predictions that still depend on an expensive and immobile physical sensor network on e.g. roads. The obtained information from passive tracking would be multimodal and available in all populated regions. This traffic data would come without history and the method therefore needs to be robust and able to work with limited data. The third part of this thesis therefore looks at ways in which current traffic prediction problem can be simplified and the method to solve it made more robust, increasing prediction accuracy.

Acknowledgements

The work for this thesis was conducted at the Department of Geography at the University of Zurich and was made possible by the support of many people that were accompanying me on my journey and helping me in countless ways. I would like to express my gratitude to all of them, as without them I could not have experienced this time of cherished personal growth at the GIUZ.

First I would like to express my gratitude to my PhD committee that guided me through the process, provided valuable suggestions and gave me a focal point every year to work towards.

I am particularly grateful for the supervision by Röbi Weibel. The freedom he gave me to explore those aspects of GI-Science that appealed most to me and in the way I saw fit was a very generous gift. This liberty encouraged me to take ownership of the whole process, from setting the goals to planning and execution, allowing me to improve in all those areas.

Second, I would like to thank Ross Purves, who always took the time when I was having a problem that he could help me with. The help and advice he provided always turned out to be on point and very valuable. For that I owe him many thanks.

In scientific terms I could benefit greatly from the late Rein Ahas from the University of Tartu, Erki Saluveer from Positium OÜ, Kay W. Axhausen, Henrik Becker and Joseph Molloy from the IVT at ETHZ and Devis Tuia, Michele Volpi, Diego Marcos, Shivangi Srivastava and Beni Kellenberger from MMRS. They enabled the research by fruitful and *very interesting* discussions and by providing the data necessary for this work.

In the office, the coffee break and lunch discussions were always very enjoyable interruptions from work, providing silliness in the sobriety of research, a heads up in more difficult times and, if I was lucky, cake. Thanks Diego, Olga, Mitch, Curdin, Pia, Flurina, Michelle, Raha, Hoda, Gilian, Meysam, Ali, Peter, Beni, Annina and Arzu.

Laying the indispensable foundations without which I would never even have thought of pursuing a PhD were of course my family and friends. A big thank you to Maya, Stufi, Lya, Christian, Bea, Stephu, Böni, Linda, Andreas, Remo, Bröni, Lukas, Ladina, Carlos, Klaus, Dorothea, Thomas (2×), Dania, Rae, David and many more of you.

Last, but definitely not least I want to thank Tschasch. Having a companion with whom to be on the awesome adventure of life is an uplifting, consoling and inspiring experience that I find hard to put into words, but for which I am extremely grateful.

Contents

List of abbreviations

ANN	Artificial Neural Network	Technique of Machine Learning
ARIMA	Auto-Regressive Integrated Moving Average	A method to analyse time series
CDR	Call Detail Records	Record produced by using telecommunication equipment
CRF	Conditional Random Field	Statistical method for structured prediction
CSD	Cellular Signalling Data	Data generated ensuring the operation of mobile phones
CV	Cross-Validation	Statistical method to estimate generalisation error
DAMOCLES	Daily Mobility Clustering Estimating Space-use	A method to interpolate sparsely sampled spatial data
DBSCAN	Density-Based Spatial Clustering of Applications with Noise	A method to identify groups of *close* points
EU	European Union	Political and Economic union of European states
FFNN	Feed-Forward Neural Network	One architecture of ANN
GDPR	General Data Protection Regulation	EU-Regulation 2016/679 concerning data privacy
GIS	Geographical Information System	Software to handle spatial data
GNSS	Global Navigation Satellite System	System allowing positioning
GPS	Global Positioning system	An instance of a GNSS
GSM	Global System for Mobile communications	Mobile phone network standard
HMM	Hidden Markov Model	A method for sequence labelling
IAPP	International Association of Privacy Professionals	A non-profit organisation in the field of data privacy
ITS	Intelligent Transportation System	System for real-time traffic management
KNN	K-Nearest Neighbours	A method for classification and regression
LSTM	Long Short Term Memory	A possible component of ANN's
LTE	Long Term Evolution	Standard for high-speed wireless communication
MAE	Mean Absolute Error	Error measure for univariate cardinal predictions
MAPE	Mean Absolute Percentage Error	Error measure for univariate cardinal predictions
NA	Not Available	Placeholder indicating a missing value
NN	Neural Network	See ANN
OD-Matrix	Origin Destination Matrix	A way of summarising movement
PeMS	Performance Measurement System	System in California measuring traffic
RF	Random Forest	Statistical method for classification and regression
RMSE	Root Mean Squared Error	Error measure for univariate cardinal predictions
SD	Standard Deviation	A measure of variability for univariate variables
SVM	Support Vector Machine	A method for classification and regression

Chapter 1

Introduction

Fun will now commence!
— Seven of Nine in 'Ashes to Ashes'

1.1 Why study mobility?

The ability – or inability – to move is a very fundamental part of the human experience. It touches nearly everything we do and it should therefore not come as a surprise that the analysis of movement can be a lens through which many aspects of the human experience can be viewed, contributing to its understanding and shaping its future.

1.1.1 Mobility studies for planning

Urbanisation today is a global trend, which, along with population growth in some parts of the world – particularly certain countries in Asia and Africa – leads to an ever increasing number of people in cities [1]. This makes cities both denser and larger and pushes transportation infrastructures to their limits.

For example, Zurich, despite its location away from the more extreme forms of rapid urbanisation in the far east, has seen an increase in mobility, as illustrated in Figure 1.1. The trend towards increased movement and ever larger number of inhabitants in the metropolitan area have been very clear, despite an interesting migratory pattern out of and back into the city proper, which will not be discussed here.

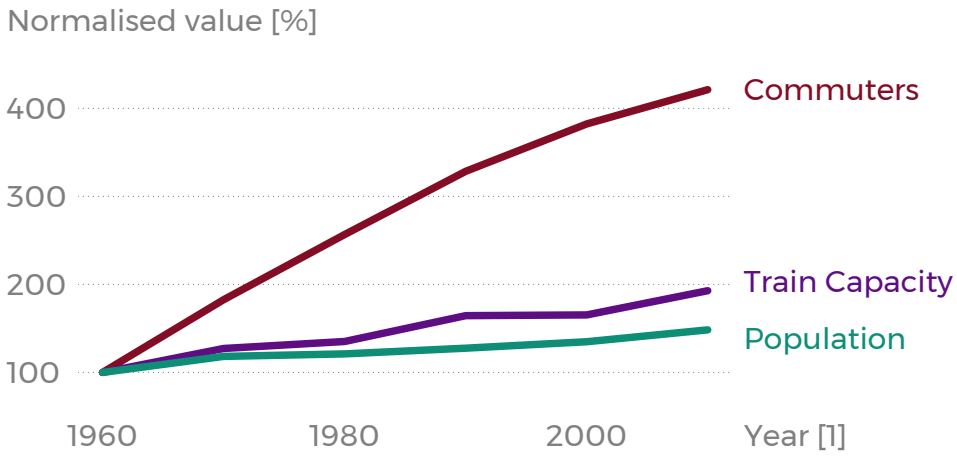


Figure 1.1:
Number of people commuting into Zurich, capacity of the public transport in the city of Zurich and population size of the greater Zurich area relative to their 1960 values. Data obtained through www.stadt-zuerich.ch

Enabling people and goods to reach their destinations quickly and comfortably through a well-designed system is the formidable task facing city planners. They have to deal with an extensive existing infrastructure that was previously built for requirements that need no longer reflect the current needs of people using the system (Givoni and Perl 2017). Furthermore, political considerations may lead to additional requests for the traffic system to not only facilitate speedy mobility but also to channel it into or out of certain geographical areas or through certain modes of transport (Gössling 2013).

[1] https://www.un.org/development/desa/publications/2018-revision-of-world-urbanization-prospects.html

To perform this task, city planners not only need significant technical expertise, but also insights into the demand side of mobility. This requires them to understand the different reasons as to *why* people move, whether this is to commute to work, deliver goods and services services or pursue leisure activities. There is also the obvious question of *when* and *where* they move, i.e. the spatial and temporal components of movement. This is necessary to assess the need for capacity between important locations. Finally, planners need an understanding of *how* people want to move in order to gauge the relative importance of transportation modes.

Obtaining data for planning

To answer these questions, traffic planners, with the help of demographers at the Federal Statistical Office in Switzerland undertake a significant effort every five years and conduct the *Mikrozensus Mobilität und Verkehr* which will be described in Section 2.1.3. This census provides a broad overview of the demand for mobility that is used to determine the need for e.g. infrastructure expansion.

However, planning based on mobility does not need to rely solely on the broad statements obtainable from the census. For example, it can also be useful to know when and where traffic jams happen in order to plan the best route and time of departure to reach a destination. This requires a much more fine grained spatial and temporal view of mobility in the region in which the movement is to take place (Kok, Hans and Schutten 2012).

1.1.2 Mobility as an economic factor

Aside from studying traffic and mobility for the purpose of planning and prediction, there is also considerable economic value in understanding such processes.

Resource allocation

One example taken from the domain of resource allocation would be a decision about where a new store of a coffee chain should be located. An easy answer combining mobility trajectories with geographic information would suggest opening it *where a large number of people travel through and the competition is tame*. However, a more nuanced understanding of mobility can significantly improve this answer.

Studying and understanding mobility means that besides the purely geometric information, knowledge on the semantics of the underlying movement and/or the individual performing that movement also becomes available. This need not be restricted to insights from movement; indeed, adding additional data onto a mobility data set is relatively easy once the individuals are identified. Having such semantic information can refine the answer above to *where a lot of commuters of a specific age and socio-economic status travel and do not have an earlier opportunity to get their morning coffee* which, economically, is much more relevant than the first statement.

Insurance

The value of mobility information can increasingly be tapped into, as customers grow more accustomed and less opposed to the idea of being tracked by companies (Tenopir et al. 2015). One example of this comes in the form of a the pay-as-you-drive and pay-how-you-drive insurance schemes (Tselentis, Yannis and Vlahogianni 2016). Insurance companies that study the mobility behaviour of their customers can better estimate the probability and magnitude of claims and thereby either offer rebates or select good risks and eschew bad ones.

From an actuarial point of view, mobility and the insights gained therefrom can be viewed as a additional feature that contributes to determining the fair premium, much like the make and motorisation of the car or the age and sex of the driver. However, from a societal point of view questions about both a surmised undermining of and an intrusion into the private sphere can be raised, as aspects of the private life of an analysed individual – such as preferences regarding how they spend a Friday evening – are revealed to a corporate entity. A person's mobility thus reveals more than the set of spatio-temporal coordinates that are recorded by a device. Those coordinates instead allow delving into the social life, as will be discussed in the next section.

1.1.3 Mobility as a proxy of the private

Home as indicator

The use of space, as alluded to in the economic examples above is correlated with many aspects of an individual's life. One of the best known manifestations of this is the difference in socio-demographic variables across different geographic locations, e.g. the postal codes of a city. Everything from age, household composition, and income to religion and country of origin is easily obtainable on a fairly granular level through open data in many cities, including Zurich [2]. Because the mobility of a person can be used to infer the location of their home, information about that person with respect to the aforementioned characteristics will be revealed indirectly through simple location updates. While on an individual level this information is probabilistic in nature and may therefore be of limited use, it is based on just a very simple extraction of the home location from movement, and therefore only represents the beginning of what is possible.

Movement as indicator

With additional effort, all places that have been visited can be identified (Furletti et al. 2013; Krumm, Rouhana and Chang 2015; Rinzivillo et al. 2014). This reveals much more precise information about the person and the semantics of their daily routines. For example, stopping at nurseries can reveal the presence of children, going to political demonstrations can reveal moral views, and attending mass on a regular basis can point to religious affiliation. While all this inference is of course uncertain – the person above could have gone to the nursery to offer a ride home to their partner (the kindergarten teacher) visited the protest as a spectator, and attended mass as a technician – it is nonetheless individualised and can have real life consequences if, for example, the government decides that participation in a demonstration is to be sanctioned.

Aggregated information

Aside from any individual impacts, semantic information on people whose mobility is known can reveal wider aspects of society at large. For example, in research conducted in Estonia, Silm and Ahas (2014) showed the diurnal pattern of ethnic segregation in the area around Tallin. The population wide pattern of nightly segregation and daily mingling was revealed by looking at the trajectories of cell phones whose operating system language was known.

[2] `data.stadt-zuerich.ch`

The study of human mobility is of interest to a wide variety of actors, be they governmental, corporate and societal. On an aggregate level, understanding mobility can help planning urban development, finding optimal routes to a destination at a specific time or decisions concerning resource allocation, while on an individual level, an analysis can reveal important aspects of a person's life.

1.2 How to study mobility

As shown in Section 1.1 there are many reasons to study mobility. The next question therefore concerns how this can be approached. There is more extensive treatment of this question in Chapter 2, so this section will only cover the broad strokes.

Interviews

Understanding mobility involves obtaining semantic information on movement. The best way to obtain semantics has been, and will probably remain for a long time, to ask people directly. Surveys and interviews have the advantage of semantic accuracy and they can deliver qualitative information that is not easily obtainable through automated and computerised methods.

This type of information acquisition is invaluable for questions about preferences, such as the type of transportation mode an individual prefers, what it would take to change that preference and so on. While not all preferences and expected responses to incentives for different transportation modes can be collected, they are a very good starting point to generate hypotheses and scenarios (Bhat 1998).

For certain types of questions however, semantic information on many individuals simultaneously is needed. For example, in order to understand the mobility behaviour of the Swiss population as a whole for example, or the visitors of a big special event, interview surveys, as accurate as they may be, have their limits. In those cases, an automated tracking system may benefit the analysis and allow for more comprehensive insights.

Active vs. passive tracking

One important distinction to make in automated systems is between active and passive tracking. In active tracking the people on which insights on their mobility are to be gained have to carry a dedicated tracker for a global navigation satellite system (GNSS) such as GPS for the duration of the study. Alternatively, they have to install a dedicated application on their cell phone that tracks their movement using the sensors that are now installed in every smartphone, including but not limited to the aforementioned GNSS sensors. In both cases, a burden is placed on the participants and on the surveyors, as they need to go through a recruitment process and have to actively do something for the data collection. The recruitment process is usually an expensive endeavour, as every single participant must be contacted individually and has to agree to participate. This problem becomes exacerbated by low response rates (Shen and Stopher 2014) which means that the effort made per recorded user on the part of the surveyor becomes even larger. Surveying techniques that ask the participants to carry devices or install and run applications incur the additional risk of manipulation errors.

In contrast, passive tracking uses data that is already being collected anyway. In particular, mobile phone signalling data (CSD) is collected by telephone companies at astonishing spatial accuracy. They are needed and used already for e.g. the localisation of a caller in case of an emergency. US-laws in that case enforce a minimal accuracy of localisation (Federal Communications Commission 2018), whereas other countries also aim for good localisations for emergency calls but have less formalised requirements.

Using this passively tracked information requires positional updates to be stored across the customer base over a certain time horizon. Even in a relatively small country such as Switzerland, a single provider generates 20 billion events or 2 Terabytes every day [3]. This deluge of data is fortunately no longer an insurmountable problem as the tools for storing and handling large amounts of data have improved significantly, and it is now possible to store all desired information about the customers. Hence, signalling data can now be used as a basis on which to perform analyses of movement.

Population scale analysis

Mobile phone penetration among adults in many so called developed countries has reached almost complete coverage. In Switzerland, one report (Newzoo 2018) puts the penetration of smartphones at 73.5%. It considers the entire population including the approximately 12% of the population below the age of 12 and makes no statements about other mobile phones leading to a conservative estimate. This means that most people can get tracked around the clock, as many nowadays rely on their phone and are compelled to take it with them wherever they go.

An additional potential source of large-scale information about space use could also come from the internet of things, in particular mobile entities such as cars. If the discussion and the hype around self-driving cars come to fruition, those cars will have to exchange information about their whereabouts amongst themselves or with a central computer. This information could then of course also be tapped into for the analysis of the flow of those cars at very fine temporal granularity and high spatial accuracy.

1.3 Research questions

This thesis explores how the goals of different groups interested in the mobility of large parts of the population can be achieved in light of the expected spread of semantically poor but abundant data with increasing spatial accuracy.

The first research question starts with a relatively commonly available data type and asks:

> **Research question 1:**
> How can the movement geometry be accurately extracted from call detail records making as few a priori assumptions on the semantic level as possible?

[3] https://ict.swisscom.ch/2015/11/from-big-data-to-smart-data-traffic-optimization-using-mobile-network-traces/

For many questions about the semantics of movement, individual trajectories of the people under study are needed. As CDR are temporally very sparse it is hypothesised that methods that work on trajectories will fail on this type of data. Instead, methods to reconstruct trajectories from CDR data that gather aggregate information over many days to reconstruct the trajectories may be able to contribute to a good reconstruction.

If the first research question started with the *data* that was available and inquired about the *methods* that can be used, the second research question flips this around and starts with the *methods* that are available for semantic enrichment and asks about the necessary properties of *data* for those methods to work:

> **Research question 2:**
> How much worse than GNSS data can passively tracked data be in terms of spatial accuracy and temporal granularity while maintaining the distinguishability of transportation modes?

Answering the first two research questions provides information about the requirements for large scale passive tracking for transportation inference. If significant parts of a population can be tracked in a way that allows traffic monitoring, applications such as short-term traffic predictions can be built without the expensive immobile infrastructure. Improving such a prediction is the focus of the third research question:

> **Research question 3:**
> How and by how much can the error in deep learning based traffic flow prediction be reduced by reframing the prediction problem and reducing its complexity?

In essence, it asks whether the good results that are obtained in other domains by using domain knowledge to state a problem in a manner that is easier to solve translate to the domain of traffic prediction. The motivation for this is that the data for traffic prediction will probably always remain limited, as traffic systems evolve and thus limit the use of old data for training. In such circumstances having an easier problem to solve will always be helpful.

1.4 Structure of the thesis

Following this introduction, Chapter 2 will provide the reader with background to the topics of this thesis. While many aspects of the theory will be covered, it will not feature a complete introduction to every aspect thereof. Pointers to the literature, however, will always be given.

Chapters 3–5, deal with the research questions in the order they are asked in Section 1.3 using real world data that will be described in those chapters.

Chapter 6 will summarise the main results obtained from answering the research questions and reflect both on the insights obtained as well as the limitations discovered while answering the research questions. In addition, on the level of the individual methods and data sources, an outlook on potential further development is provided, reflecting expectations into what direction research efforts could be undertaken.

Chapter 7 finally concludes by summarising what the contributions of this thesis were and how the research questions were answered on a very high level. Keeping with this broad overview, an outlook on the field as a whole concludes this thesis.

Chapter 2

Background and theory

Someone's sitting in the shade today because someone planted a tree a long time ago.
— Warren Buffet

2.1 Human movement

This section provides a brief overview of how human movement can be reasoned about in the context of computational analysis. First, the spatial and temporal scale of movement that this thesis is concerned about is delineated against smaller and larger scales that are fascinating in their own right and studied in great detail elsewhere. After fixing the scale, different ways in which movement can be abstracted to be computationally tractable will be discussed. It is necessary to abstract human movement in order to sensibly discuss it using computer tools. The kind of abstraction used determines what kind of questions can be asked of and answered by the data. Some relevant sources of data that can be tapped into will then be briefly discussed and compared. Finally, given that data on movement is potentially very sensitive, this section ends with a brief introduction to the data privacy regulations relevant to this thesis.

2.1.1 Scales of human movement

Human movement can be addressed from multiple angles and capturing its essence is a highly philosophical issue. While creating a complete ontology of movement would constitute a highly stimulating exercise, it is not undertaken by many researchers in computational analysis of human movement.

Micro-scale

Beginning from the smallest scale of movement there are (partial) ontologies of movement that originate in application domains such as computer vision (Saad et al. 2012) or dance (El Raheb and Ioannidis 2011). The goal at this scale is to be able to reason about the movement of the body more or less relative to itself and for example infer poses or activities such as getting up or swinging a tennis racket from simplified representations of the human, such as those presented in Figure 2.1. The reason ontologies are developed is that they provide a comprehensive universe of the poses and activities that can be talked about. In addition, they also allow for a formalised representation, enabling for example searches, as their typically hierarchical nature permits the concept of subclasses. In this fashion, atomic movements can be combined into a composite movement that carries more meaning (Saad et al. 2012).

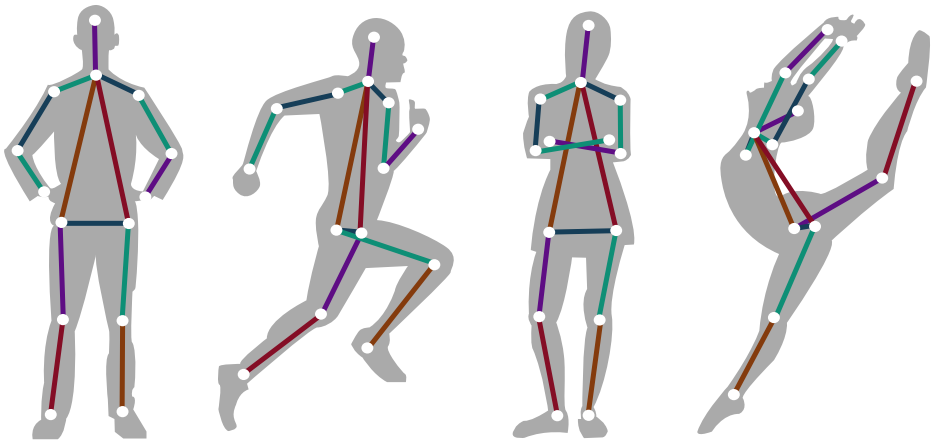


Figure 2.1:
Simplified representations of human limbs in different poses. Their relative orientation and/or movement can be used to infer the semantics of a pose or activity of a person. This scale of movement is finer than the scale at which movement is discussed in this thesis. Figure adapted from Bearman and Dong (2015).

In practice, however, at least in computer vision, instead of a complete ontology only a predefined set of activity labels are often used, between which the computer systems are tasked to distinguish, and that set typically reflects the source of the data. In a review by Poppe (2010), the sets of activities to be detected barely overlap and include very specific activities such as *sideways gallop, throw from bottom up*, and *swinging a baseball bat and walking*. In the sets of activities presented in a survey by Lara and Labrador (2013) on the other hand one can find both *standing still* and *riding escalator*, which, to the person performing those activities, may feel the same but are semantically different because of the context. Due to the different angles from which movement is analysed there has not yet been an all-encompassing ontology capturing all nuances for the different fields that could readily be used.

Meso-scale

When changing the focus to a larger, geographical, spatial scale, the importance of the motion of the human body with respect to itself fades, to be replaced by changes in the position of the body as a whole. This geographic scale brings with it a coarser grained temporal scale, as the time required to move a noticeable distance usually exceeds the time required to *get up* or perform other typical activities on the finer scale discussed previously. Although the duration of the micro-scale activities may well exceed the duration of movement on the meso-scale, the relevant scale at which the semantics are inferred is always finer on the micro-scale than on the meso-scale.

At any given time, the position of the body as a whole is often represented by a single point, because in analyses at that scale, the orientation and extent of the body is usually irrelevant. These points can then be combined into lines, annotated with relevant labels, such as mode of transportation and set into a relevant geographical context, such as proximity to public transport stations or roads, as shown for example in Figure 2.2. As this is the primary scale used in this thesis, its implications will be discussed in detail in later sections.

Figure 2.2:
Example trajectory of meso-scale movement in terms of orders of magnitude for distance and time. In this example the data quality is relatively good and the beginning and end of the bus stage are clearly close to bus stops, conforming to expectations.

Macro-scale

Moving to the limits of the geographical scale in terms of space and time, one can also study migration patterns, as shown by Lu et al. (2016). On that scale, even the daily patterns of movement become less relevant, as the relevant granularity can be based on broad political boundaries larger than typical daily ranges of motion.

Figure 2.3 depicts macro-scale migration patterns from Africa and the middle east towards Europe. These movement patterns are currently the focus of lively political debates (Geddes and Scholten 2016) and has found its way into party programmes all over Europe [1]. Treating them would require a different set of tools, as the experience of this kind of mobility may entail so much more than just the transportation modes that are of concern here. Also the geography in which the movement happens is vastly different from the Balkans to the Sahara desert. Those differences are interesting and worth studying in their own right. However, this is not in the scope of this thesis.

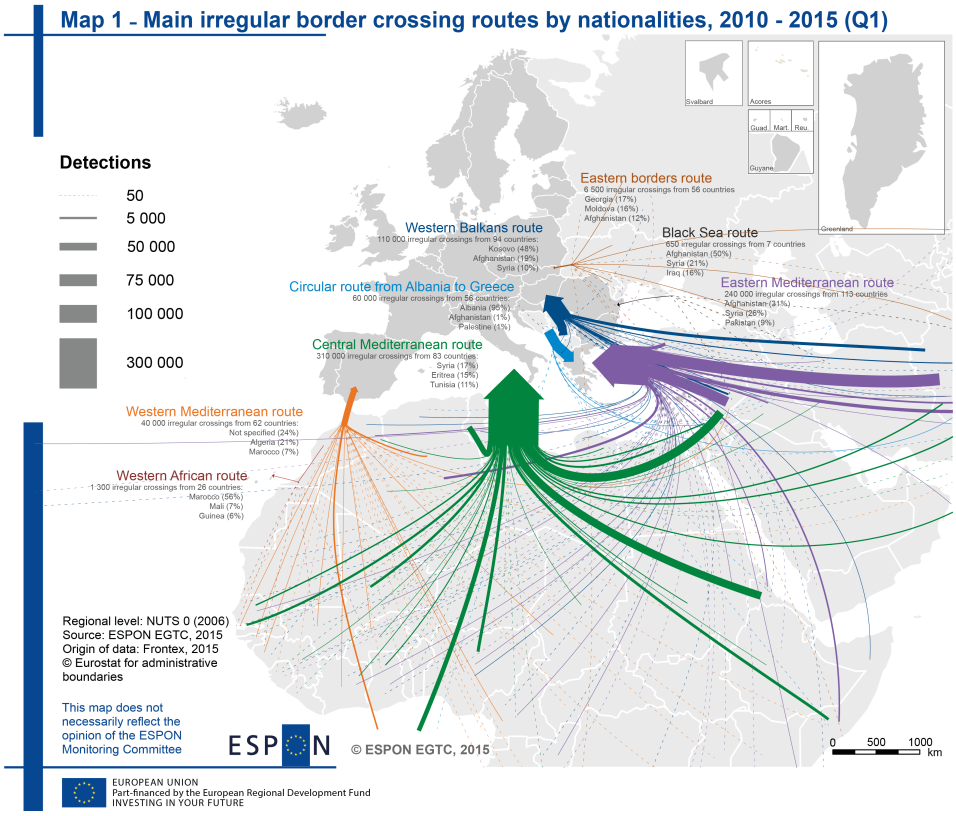


Figure 2.3:
Aggregated information on movement on the macro-scale. These trajectories are out of scope of this thesis, both in terms of the duration of the trajectories and in terms of distance travelled, despite its relevance for the political discourse. Image was sourced from the European Observation Network for Territorial Development and Cohesion [2].

In this thesis the focus lies on what could be termend *the daily experience* of movement in western countries. While longitudinal data are used, the focus is on the movements that people could refer to or think of in an answer to the question "Where did you go today?".

[1] https://www.afd.de/zuwanderung-asyl/, https://www.ukip.org/ukip-manifesto-item.php?cat_id=5, https://sd.se/vad-vi-vill/, https://www.rassemblementnational.fr/pdf/144-engagements.pdf, https://www.svp.ch/partei/positionen/themen/asylpolitik/

[2] https://www.espon.eu/

2.1.2 Abstractions of human movement

Having established the scale of the movement to be discussed in this thesis, the language, i.e. the abstraction in which it is discussed needs to be specified. This language defines and thus limits the entities that can be talked about and the usefulness of data can depend heavily not only on the precise questions that are being asked but also on the language in which it is expressed.

While the different abstractions of movement presented here obstruct the view on some aspects of movement, researchers using them typically do not go so far as to claim that their abstraction describes how and what movement *is*. In fact, the main concern of those analysing movement data is seldom the nature of movement as such and therefore they accept the expressive limitations of their abstractions of movement as long as it does not infringe on their ability to perform the analysis at hand.

While there are many different abstractions of movement, only those relevant to this thesis will be discussed here.

Geometric abstraction

The geometric view on movement in the Euclidean space of GNSS data considers sequences of points that contain a spatial and a temporal component (Laube 2014). While other representations of movement, especially for multiple individuals are possible (cf. below), this representation is presented first to match the order in which they appear in the thesis. In its simplest form – which forms at least part of most analyses – the points that build this sequence take the form of tuples (id, x, y, t), where id determines the identity of the tracked human or sensor, x and y determine the point in space and t determines a point in time where the moving body has been (Das, Ronald and Winter 2014).

Every such row (or vector) containing at least the four elements (id, x, y, t) is called a fix, as it ties the (estimated) position of the moving person to one precise point at a given time. There may be more elements to those vectors that were either measured by sensors (Gong et al. 2012) or added by the analyst based on data (Stenneth et al. 2011), so the definition above only states the minimal requirements.

Fix

If this information pertains to a single individual and is ordered by time, the set of fixes is called a *trajectory*. Defined this way, a trajectory is not the continuous actual movement itself, but a potentially imprecise and temporally discretised representation thereof. Furthermore, the term is used in this thesis to denote any sequence of measurements that contains at least the four mentioned variables, even if there are additional ones that contain further information. It is important to note that while movement happens in three dimensional space, for the purpose of this thesis, using this definition that uses projections on a two dimensional space is sufficient.

Trajectory

It is obvious, a human body does not record those positional updates, and so the recorded position is actually that of a recording device that the tracked person is carrying. This directly impedes analyses at the micro-scale as the movements of the different body parts relative to one another cannot be recorded this way. Furthermore, positioning is never truly accurate (Ranacher et al. 2016), with positional uncertainties typically being significantly larger than the extent of the human body .

Nonetheless, in many cases the recorded positional information is typically equated with the *point position* of the person to be tracked. The mismatch between the position of the device and the person is simply too small to pose any practical problems and the theoretical issue with the ill-definedness of *the point position* of a human body with spatial extent is purely of academic interest at the geographical scale.

Recording errors

The problem of uncertainty regarding the positional estimates that constitute the data can, however, be a real concern. If it is, this uncertainty is typically addressed by preprocessing the data.

One way of dealing with this problem is to use a rule based procedure that draws on domain knowledge to detect and remove "unreasonable" parts of the data (Lv, Chen and Chen 2012; Umair et al. 2014), mostly to eliminate gross outliers.

But even if there are no enormous outliers, using the raw data can be problematic. As Ranacher et al. (2016) have shown, deriving features from the raw trajectory can be erroneous already under very mild assumptions on the error process, especially if the temporal granularity of the collection is very fine. Therefore, often smoothing is applied. This is an activity that replaces the raw trajectory with one that broadly follows the one that was measured, but shows less volatility, for instance in terms of turning angle or velocity. The hope is that the smoothed trajectory will be a better representation of the actual movement than the measured raw trajectory.

Smoothing techniques can assume a distribution of the errors on the position estimates and use that knowledge to obtain an updated estimate that can be expected to be closer to the actual position of the recording device (Horn et al. 2014; Lin, Hsu and Lee 2013). Simpler methods such as kernel smoothers can also be used (Schuessler and Axhausen 2009) These two preprocessing steps may also be combined (Kerr, Duncan and Schipperjin 2011) to obtain smooth trajectories free of outliers. Whichever method is chosen, care has to be taken in order to avoid enforcing a priori assumptions about movement onto data which do not conform to those assumptions. For instance, the trajectory of a drunk person may really be non-smooth and sudden stops may happen in an otherwise smooth trajectory, both of which may be smoothed away if one is not careful.

However, all the techniques to clean the data do not affect their shape. While their properties may have changed and their representation of the underlying movement may have improved, ultimately they will still look largely like a list of entries in the form (id, x, y, t).

Applications

The geometric representation of movement lends itself naturally to questions that compare locations and times. One of the very famous examples of this is Hägerstrands time geography (Haegerstrand 1970; Lenntorp 1999) that, given a point, reasons about which future combinations of coordinates are possible. Other examples include all kinds of accessibility studies (Neutens 2015; Tenkanen et al. 2016). Accessibility studies reason not only about one trajectory but about the interplay of several trajectories, even if the objects to which accessibility is assessed do not move over time. This restriction is not necessary and even the interplay between multiple moving objects or persons can be analysed (Stenneth et al. 2011).

Movement captured in this way is highly specific to the tracked individual and it is very rare that two individuals exhibit nearly identical trajectories over extended periods of time. Therefore, given as few as four fixes of a person, it can be possible to match an ID in a pseudonomised dataset to that individual (De Montjoye et al. 2013). This has implications for data protection requirements, as discussed in Section 2.1.4.

Eulerian vs. Lagrangian movement

Depending on the problem to be solved, individual trajectories may not be required. Instead, simply knowing the aggregated volumes of people at certain points in time and over certain spans of time may be sufficient. In that case, one has what is called *Eulerian* Data, as opposed to *Lagrangian* data, which is how the trajectory type of movement data is also known. The names of the two types originate from alternative specifications of the flow field, but have since been applied to the movement of people (Laube 2014).

In Eulerian data, the identified entities are some form of counting stations which measure how many individuals pass by them. This information is often aggregated over time windows (such as 15 minute intervals) and only the aggregated results are made available. The data itself therefore no longer directly contains information about any individual. To also ensure that no information about any individual can be deduced, some additional precautionary measures may be necessary, such as masking information if the number of individuals in a time interval is small enough and/or randomly distorting the numbers (Culnane, Rubinstein and Teague 2017).

Thus, individual trips and derived insights, such as OD matrices are no longer calculable, as only the total numbers at the origins and destinations are known, but not how to link them. Thus Eulerian data are considerably less problematic from a privacy point of view and can – given the necessary precautionary measures – be shared or published. However their use is limited to problems where knowledge of individual trajectories is not required. Such problems include the partition of cities into functional regions and comparison between the studied cities (Louail et al. 2014), general statements on the dynamics of humans within cities (Candia et al. 2007) or deriving a "geography of human activity" (Reades, Calabrese and Ratti 2009).

While the privacy preserving properties of this kind of data is conceptually enticing, the increased capacity to store data on the individual level has led to many more recent works in the field to be based on Lagrangian data. This is a consequence of "a world that is fast becoming digital in all its dimensions" Batty (2012) allowing for different methods of analysis combining different data sources that may lead to more insights as demonstrated by, amongst others, Toole et al. (2015).

In summary, geometric representations of movement are a natural choice when reasoning about the *when* and *where* of movement, both in isolation and in relationship with other (static or mobile) objects. This information allows for easy identification and can thus be problematic from a privacy point of view. The Eulerian representation of movement data can be one approach to mitigate this problem, maintaining privacy if certain preconditions are met. However this comes at the cost of reduced utility of this kind of data.

Semantic abstraction

Not everyone is interested in the pure geometry of movement through time-space at a geographical scale. As indicated in Section 1.1, many relevant questions revolve around *why* and *how* people move the way they do. This requires information that is not contained in the geometry of the movement alone.

The questions in this domain are diverse and as a consequence so are the ways in which researchers talk about movement on a semantic level. For the purpose of illustration and because it is relevant to this thesis, only an example from traffic planning will be given here.

Semantic partitions

In traffic planning, the semantics is mostly concerned about the *how* of movement. A frequently appearing taxonomy of movement is that of *stages*, *trips* and *tours* in combination with *activities* at which *trip purposes* can be fulfilled (Axhausen et al. 2003). Because the context is given by the transportation domain, the concepts are informed by its semantics.

The primary units for describing movement are the *activities*. These are what motivates people to move in the first place, as certain activities are only possible at certain locations. Although in this thesis the activities are not of interest, they do mark the beginning and end of movement and are thus relevant.

The movement between two activities is called a *trip*. Trips may be very diverse in all possible aspects such as duration, used transportation modes and so on, but no trip is interrupted by an activity. Instead every trip has an activity both preceding and succeeding it.

A maximal, contiguous part of a trip during which the same mode of transportation is used is called a *stage* – for example the part of the movement from where a person enters a bus until they leave it again. This is the atomic semantic unit in which movement becomes partitioned and any finer distinctions such as movement within the coach, sitting down on a seat and so on get disregarded.

Different trips can be grouped together to form a tour, provided the location of the activity preceding first trip coincides with the last trip's succeeding activity's location. If a tour is part of an even larger tour one can call the smaller one a subtour.

While this description of movement seems intuitive, there are grey areas that do not always get treated in the same way. For example, in some studies there is a *stationary* class that can capture transition times between modes of transportation (Bantis and Haworth 2017), while others could only speak about transitions using a *walk* stage as all possible classes represent actual movement (Bolbol et al. 2012). While such a distinction is not needed for all analyses of movement data, they can be of importance to transportation planners when, for instance, optimising schedules. Other semantic grey areas could revolve around when an activity is split: Does a bathroom break already constitute a trip and a new activity? Can the same be said for getting up from the chair to get a book from the bookshelf? Therefore in every application, the specific requirements will determine the details as to how exactly the terms get delineated.

The partitioning of the trajectory into stages is only possible because the semantic layer provides the positions of the breaks. If this layer is not available, i.e. on the geometric level, movement is often only partitioned into *moving* and *staying* segments, as argued by Parent et al. (2013) and the mapping between the semantic terms, which are of interest to the analyst, and the geometric terms, which are easily readable from the geometry are by no means trivial.

Geometric vs. Semantic partitions

The activities of the semantic taxonomy can broadly be matched to "long" stays of the geometric taxonomy, begging the question of the definition of "long". While a trip can correspond to a single move segment, waiting during a change in modes of transportation or stopping for a red light can already introduce a new (geometric) stay segment without interrupting the (semantic) trip. On the other hand in some cases the mode of transportation can be changed without ever fully stopping (walk to run, walk to tram, walk to bike) resulting in move segments comprising several stages. Thus neither partition of the movement is a refinement of the other. This reflects that they describe different aspects of the underlying movement.

The motivation for movement – the *why* – is given by the *trip purpose* which is usually assumed to be found in an exhaustive list (Bohte and Maat 2009; Jiang et al. 2013a; Wolf, Guensler and Bachman 2001). Trip purposes are usually assumed to be tied to geometric stay locations. The lists of the purposes are typically short in comparison with the diversity of people's lives. Thus, the level of detail that can be captured with such a typology is limited, but apparently often considered sufficient. Sometimes the lists include a class for all purposes that do not match the categories (Bohte and Maat 2009) which can be seen as a concession to the complexities of the human experience. Naturally there is a wide range of possibilities to infer those trip purposes, as presented in the review by Gong et al. (2014).

Having combined geometric and semantic information on the trips undertaken by individuals, practitioners and scholars can use them to understand a wide range of transportation related issues, such as the impact of neighbourhood composition on mobility (Rutherford, McCormack and Wilkinson 1996).

Need for semantic enrichment

Knowing the semantics of movement allows for important insights that cannot be derived from the geometric aspects of movement alone. However, as will be discussed in Section 2.1.3, purely geometric data are easier to obtain in large quantities than data with carefully curated semantic information. This has led to a situation in which there there is a lot of data, that cannot directly be analysed semantically, as the semantics are often missing and have to be inferred first.

One way to remedy this lack is by obtaining estimates for the semantics of movement from its geometry. This has been done for the tasks of identifying trips, inferring mode of transportation and imputing trip purpose (Bohte and Maat 2009; Jiang et al. 2013a; Parent et al. 2013; Shen and Stopher 2014).

In addition to the pure geometry of movement, there is a semantic layer that often lies at the core of what is relevant to the analysis. Unfortunately the semantic level is harder to obtain than the geometric one and often the latter has to be inferred from the former. This enrichment is complicated by the fact that the semantic partition of interest need not match the partitions that are available from the geometry.

2.1.3 Data sources

In the literature, a wide range of information sources can be found to generate insights into human mobility at the meso-scale. In this section those relevant to this thesis, the basics of how they work, the shape of data they produce, their results, and their implications for studies that use them will be discussed.

Surveys and questionnaires

Transportation scientists and planners have long been interested in why, how, where, and when people move. For the most part, they have fulfilled their information needs by directly asking people about their travel behaviour. Stopher and Greaves (2007) provide a useful overview of the history and state of transportation surveys in 2007, incidentally the year in which the first iPhone was introduced, which ushered in the profound changes that have since been seen.

Content of surveys

According to the authors, a survey is performed by asking a carefully sampled, relatively small (Typically no more than 3%) fraction of the population about their movements. The results would comprise a sequence of activities and trips that connected those activities. Each activity can have information about the motivation behind it, the duration, and so on whereas the information about the trips would comprise duration, mode of transport, and potentially other factors. The sequence typically comprises all trips and activities undertaken in one full day, although there are notable exceptions that spanned longer durations (Axhausen et al. 2002) .

The respondents would be contacted either through telephone calls or in written form. This requires a rather significant logistic effort on the part of the surveyor, especially because even just 1% of a large population (such as an entire country) may be a lot of people and not all contacted potential participants end up delivering usable data. Stopher and Greaves (2007) cite an initial response rate of approximately 60% for a good north American survey by telephone, of which another 60% then complete the survey, resulting in an overall success rate of only 36%.

Swiss survey

In Switzerland, the biggest such survey is known as *Mikrozensus Mobilität* and is conducted once every five years with about 60000 participants (Die Schweizerische Eidgenossenschaft 2018). The response rate was 53% in 2015, thus over 100 000 people had to be contacted in order to obtain the desired number of participants.

The 60 000 people that are surveyed in Switzerland make up a very large number, unquestionably resulting in a considerable effort for the surveyors. Yet compared to the whole population it is still relatively small at only 0.77%. While the people are carefully selected so that all regions and ages are represented, there is still an uncertainty incurred by sampling. Furthermore, the reference days on which the participants were interviewed were chosen randomly, the effective fraction of observed people on any given day is therefore negligible with less than 200 people on average.

In this survey, all stages longer than 25m outside the home have to be reported as well as beginning and end times, mode of transport, and purpose of the trip. Additional questions are also asked in relation to personal circumstances such as ownership of vehicles and public transportation cards, as well as the composition of the household as a whole. This yields as output very detailed information about the person interviewed with the (hopefully) complete information about the mobility of one day.

Limitations

The direct contact necessary for such surveys has the obvious advantage of accuracy in the sense that the semantics of the movement are likely to be correct. However, traditional surveys are known to have a problem with under-reporting (Pereira et al. 2013; Shen and Stopher 2014; Stopher and Greaves 2007). Further problems originate from an uneven response rate between the different strata of the population, forgotten trips, the relatively high burden on the participants, and the price of obtaining the data which grows linearly with the sample size. (Bricka and Bhat 2006; Bricka et al. 2009; Furletti et al. 2013). This means that while the information on those trips that are reported are likely to be good, there can be a rather significant number of trips that do not even make it into the records and the trips can, despite all efforts, not be sampled evenly across the population, negatively affecting the analysis.

Furthermore, there are the drawbacks of scope and temporal granularity, as illustrated by the case of Switzerland [3]: The surveys happen in very long intervals. In Switzerland, as mentioned, the time between two surveys is 5 years, during which the traffic cannot be assumed to remain (nearly) constant throughout the country. As seasonal effects such as a particularly warm or rainy summer can impact the observed results, sampling at such low temporal rates incurs high variability, even in the case of perfect coverage.

The surveys that are performed therefore amount to looking at many disconnected individuals with a magnifying glass, enabling analyses that are bound to be based on very generously aggregated data, as for example done on the official FSO report [4]. In terms of how the population as a whole thinks – and more importantly: acts – on mobility, this is a fantastic source. The limitation it has lies in its lack of coverage of traffic situations in specific regions at specific times, where the information is insufficient.

Surveys and questionnaires are an excellent source of semantic information if the attitude of a population to traffic and the relationships between preferences, measurable covariates and mobility behaviour is to be investigated. Also they often represent the most detailed information about the general mobility behaviour of individuals available. However, due to their limited scope and long intervals between two consecutive large surveys, they are not suited to measure traffic at a given time and location.

Loop detectors

For certain specific questions, having full knowledge of the travel behaviour of (a representative sample of) the population may not be necessary. Instead, it may suffice to know about certain aspects of the network of a single traffic mode. One mode for which many questions are of that nature is that of motorised vehicles on the street network.

Functionality

The way most administrations go about this is by using so called loop detectors. The technology itself dates back to the 1960's and Anderson (1970) provides a overview of how they work which will be summarised here. Usually loop detectors are based on the following principle: Parallel to the street surface, there is a (mostly rectangular) induction loop through which a current flows, generating a magnetic field above the loop. Any vehicle that drives over such a loop will change the self-inductance of the loop, creating a measurable signal. Such loops can thus measure the number of times the loop changes from *occupied* – i.e. there is a car above it – to *non-occupied* and back, thus yielding the number of vehicles that have passed it. Additionally they can measure the total time during which they were occupied, the so called *occupancy rate*.

[3] https://www.bfs.admin.ch/bfs/de/home/statistiken/mobilitaet-verkehr/erhebungen/mzmv.assetdetail.4262242.html

[4] https://www.bfs.admin.ch/bfs/de/home/aktuell/neue-veroeffentlichungen.gnpdetail.2017-0076.html

If a second such induction loop is placed in close proximity to the first, additional measurements such as speed and length of passing vehicles can be measured (Wang and Nihan 2003). Such a system is called a *dual loop*. While some systems use heuristics to calculate traffic speed from single loop detectors (PeMS 2017), measuring speeds and vehicle lengths directly is clearly preferable.

The denser the loops are positioned, the more fine grained the information can be, especially if the temporal granularity is very fine. Too fine a resolution in time and space can be a potential point of a re-identification attack on users' privacy based on this Eulerian data.

Compared to travel surveys, data from this source have different properties. The advantage of such systems is that the burden on the people whose data are being collected is zero. Drivers on roads that have loop detectors do not even notice that their data are being collected. Furthermore, data on *all* vehicles that drive over those detectors are collected, providing for a complete picture of traffic at the location of the loop. The downside for this large scope of collection is the narrow focus on a single mode of transport (individual motor car traffic) as well as a loss of granularity, since individuals can no longer be tracked.

Situation in Zurich

In the canton of Zurich there are 273 measuring stations [5]. However, the data are unfortunately not available through an API but through individual PDF's per station (if at all) and are already averaged over times of day or days of the week, effectively making any detailed analysis impossible for people outside the authority.

Loop detectors provide Eulerian data on a single mode of transportation and provide coverage of all vehicles that drive over it at relatively fine temporal granularity and can be shared in a privacy preserving manner.

Public transport measurements

Conceptually similar systems to loop detectors in roads exist for public transport. Closest in spirit are simple measurements on occupancies of vehicles, using technologies such as cameras (Chen et al. 2008) or infrared sensors (Gerland and Sutter 1999). For these systems, the same advantages and limitations apply as for the loop detectors.

[5] https://afv.zh.ch/internet/volkswirtschaftsdirektion/afv/de/verkehrsgrundlagen/instrumente_und_erhebungen/verkehrszaehldaten_kanton_zuerich.html

In addition to measuring counts on the vehicles, certain public transport systems require the traveller to check-in when they start using the transportation and check-out when they have reached their destination (Liu, Biderman and Ratti 2009), typically using smart cards. The properties of this kind of data comprise elements from both types of information presented thus far: They allow following the individual, as smart card IDs are known. In addition, all passengers are recorded without additional burden to the passengers, as checking in is a prerequisite for using the service. But again, the scope of data collection is limited to the service(s) using the smart card system and any trip outside the system (using for example bikes or private cars) will not be captured.

GNSS trackers and cell phone apps

Smart card systems have overcome the passenger systems' shortcoming of not being able provide data at the granularity of the individual passenger. However, they still suffer from their restriction in terms of traffic modes and reach (if there is a significant fraction of non-customers). A system that aims to overcome this shortcoming must therefore either be omnipresent in public space or closer to the people on whom data are to be gathered. One way of moving closer to the traveller is by tracking them with a sensor. Using Global Navigation Satellite System (GNSS) technologies such as GPS, this can now be done with sub-metre accuracy (Cai et al. 2015), allowing – in theory – for very detailed analyses. In practice however the spatial accuracy of the fixes is often worse, especially in cars or trains but also in urban canyons (Modsching, Kramer and Hagen 2006).

Phones and pure GNSS sensors

While GNSS studies in the early stages of the technology required external devices (Wolf et al. 1999), most studies now rely on GNSS sensors that are built into smartphones most people own in many western countries (Federal Statistical Office 2018).

Using the mobile phone rather than a dedicated GNSS sensor offers some advantages in terms of tracking mobility. First, the mobile phone can make use of non-satellite signals to position itself, as is done for assisted GNSS or WiFi signals (Zandbergen 2009). Thus, even in situations where GNSS reception may be patchy, such as underground tunnels or in trains, some positioning information may be available, because there is usually reception of mobile phone signals. This leads to wider coverage of mobile phone tracking when compared to tracking with dedicated GNSS loggers, even if the latter are usually of superior quality for estimating positions purely based on GNSS signals.

Furthermore there are other sensors available to the Smartphone that can provide further information on the tracked person, such as accelerometers that can provide useful information for traffic mode detection, either on their own (Hemminki, Nurmi and Tarkoma 2013) or in combination with GNSS signals (Nitsche et al. 2014). Transport researchers therefore have been following trends in the use of GNSS sensors very intently, as they could help to bring down the costs of their surveys while yielding good results compared to self-reported data (Shen and Stopher 2014), albeit still not perfect (Vij and Shankari 2015).

Mobile phones or dedicated GNSS trackers both record positions continuously. Consequently, they do not suffer the problem traditional surveys have that trips can get forgotten (Nitsche et al. 2012). The burden placed on the participants of handling and carrying a recording device is usually considered lower than the one of a lengthy telephone interview. The participants have to deal with a device, but the interaction can be limited to charging and carrying the devices around in a pocket during the day, both of which happen nearly automatically if smart phones are used. The continuous tracking of people also constitutes an improvement over infrastructure-based collection methods because they are not bound to a single mode of transportation and thus can be used for multimodal transportation mode detection. The fact that this approach to obtaining data requires recruitment means that through direct contact communication beyond the pure location signals is still possible, allowing for the detailed semantic information associated with traditional surveys and an awareness of the tracked people that their movement is being recorded (although explicit consent may be obtained otherwise).

Added value and limitations

With GNSS, more longitudinal analyses are possible than with surveys, allowing the observation of repetitive and habitual behaviour. This in turn opens the door to deeper insights into a person's life. This way not only transportation mode detection becomes possible but also the detection of all significant places in people's daily lives, (Zhou et al. 2007) and the analysis of their geo-social behaviours (Farrahi and Gatica-Perez 2010).

However, both the use of GNSS sensors and mobile phones as sensors entails direct contact of the surveyor with every single individual being tracked. This ensures that the explicit consent is actually given with full understanding of the consequences, which may not the case if it was given as part of an end user licence agreement or similar. On the other hand the need for direct contact also limits the scope of the analysis, as the effort and thus cost of such a survey grows with each tracked individual. As a consequence, while these techniques can potentially replace the travel survey, the sample sizes will not be large enough to replace systems like loop detectors for traffic monitoring.

GNSS like data obtained through dedicated trackers or mobile phones offer rich geometric information that can be enriched semantically through inference, enabling deeper insights into the tracked people's mobilities and by extension: lives. However, as the effort to collect this data scales linearly with the number of users, it will hardly be possible to design a survey with significant coverage of the population if it does not include access to a large database of already collected data.

Passive mobile phone tracking

Having the sensor close to the people whose mobility is of interest allows for multimodal observations while reducing the burden on participants. Passive tracking could, in addition, do away with the necessity for interaction between those observed and the surveyor, at least from a technical point of view. Telephone companies now have the capability to estimate positions of people relatively accurately based on the communication between cell phones and the antennae of the mobile phone system.

The most purely passive tracking approach that is possible within the mobile standard LTE is U-TDOA, or uplink time difference of arrival, where no action on the end of the user equipment (the mobile phone) is needed and the position is determined on the side of the network alone (Hamdy and Mawjoud 2012). For this, the distances to the user equipment to the different antennae receiving a signal are calculated and then the technique of multilateration is used to determine the position. Thus this kind of positioning even works when positioning services on mobile phones are disabled .

While not as precise as data from GNSS sensors, this kind of data can nonetheless be used to infer aspects of the mobility of the people in the system, (Ahas et al. 2008b; Swisscom 2018; Widhalm et al. 2015) or of the population at large (Ahas et al. 2015; Blondel, Decuyper and Krings 2015; Doyle et al. 2014; Eagle, De Montjoye and Bettencourt 2009; Steenbruggen, Tranos and Nijkamp 2015; Trasarti et al. 2015).

CDR

There are several types of passively tracked spatial information. The first are call detail records (CDR). They are generated in the operation of a mobile phone network and contain all relevant information about the communication (i.e. calls and text messages) (Wang et al. 2010). In the part of this thesis that used this type of information, this included the timestamp of the activity, the type of the activity (text message, call), the ID of the cell in which the activity took place (i.e. the antenna/mast that was used to transmit the signal), and the result of the activity (was the call incoming or outgoing, and, in the latter case, was it received, occupied, or rejected). These records are relevant for billing purposes and therefore have to be stored to ensure continuing operation of the business. The reason why this is relevant in the context of this thesis is that the cell ID, which is part of the CDR, carries implicit geographical information that can be made explicit if a database containing locations of cells is used. The following elaborations on cell based localisation are taken from Trevisani and Vitaletti (2004) if not otherwise specified. The location of the mast transmitting the signal is known and can be used as a proxy for the position of the mobile phone. Although crude, this is still used in parts of Europe to localise emergency calls, at least as of 2014 (European Emergency Number Association 2011), whereas other countries have since switched to the advanced mobile location (European Emergency Number Association 2018).

The advantage of this simple approach is that it is cheap, as it does not require any technology or systems that are not already in place. Its main disadvantage on the other hand is its imprecision. Because mobile cells can have ranges up to tens of kilometres, the information on the cell ID can be of limited use.

Using additional technology either on the side of the mobile phone – measuring time differences of signals from masts in combination with their known positions – or the masts themselves – using multilateration of the mobile phone's signal – the problem of crude location updates can be at least somewhat mitigated.

However, this requires effort either on the side of mobile phone producers or the network operators and cannot be taken for granted, although positioning features prominently in the mobile communication standard LTE after release 9[6] and promises even better data in the near future (Dammann, Raulefs and Zhang 2015). This means that over time increasingly accurate information should be available on an ever increasing number of people.

A second disadvantage of CDR's is that they are distributed very unevenly across the population (Gonzalez, Hidalgo and Barabási 2008) and can be temporally very sparse (Ranjan et al. 2012; Schulz, Bothe and Korner 2012), necessitating techniques other than those applied on trajectories to obtain insights. One such technique will be discussed in Chapter 3.

CSD

Instead of finding increasingly clever ways to deal with the limitations of CDR, an alternative is to use cell signalling data (CSD) instead. This data uses all communication between the mobile phones and the transmitting infrastructure, not just calls and text messages (Janecek et al. 2015). Those signals are generated if either the device is active, i.e. produces CDR or transmits data through wireless internet, or if the device has to establish contact with a base station because the previous one is no longer available. In the context of movement analysis this is immensely useful, as one of the reasons why base stations are no longer available is if the phone moves outside the area served by the station, i.e. if the person carrying the phone is moving. These data get generated at a much higher frequency than CDR, even when the phone is not active and most often when the phone is being moved.

Impact

If the information gathered from those kinds of systems can be made accurate enough for tasks that are hitherto mostly performed using GNSS or similar data from smart phones, passive tracking data could combine the (almost) population wide scale of infrastructure-based tracking systems with the multimodal scope afforded by GNSS technology and telephone surveys and become an addition or replacement for some of those systems. Many of the insights available today only through the *Mikrozensus* could be made available from this kind of analysis not only based on a much larger sample, but also in almost arbitrary temporal intervals. This would allow for longitudinal analyses that today are simply not available at that scale.

While for a mobility researcher such a data source must appear very exciting, to people worried about data privacy it must appear like a totalitarian nightmare. However, the GDPR which came into effect recently and the new Federal Act on Data Privacy in Switzerland which is currently in the legislative process place clear limits on how personal information can be processed and some of these limits are presented in the next section.

[6] http://www.3gpp.org/specifications/releases/71-release-9

Passive tracking can be spatially less accurate and temporally less fine-grained than GNSS based positioning, but potentially available at very large sample sizes and over very long periods. If the accuracy of passive tracking data is high enough for tasks typically associated with GNSS based analyses, they could have a profound impact on the way human mobility is studied.

2.1.4 Data privacy

Given the detailed and semantically rich information that can be inferred from positioning data – correctly or otherwise – access to them is regulated and so the obligation to handle such data and the conclusions derived from them with utmost care is not merely a moral one. This section contains a brief summary of the main points regarding Swiss and EU regulations on data privacy as these pertain to the analysis of human movement based on real data.

Switzerland

In Switzerland there are two main regulations to consider when dealing with personal data: The *Federal Act on Data Prodection* (FADP) (Swiss Confederation 2014) and the *Ordinance to the Federal Act on Data Protection* (DPO) (The Federal Authorities of the Swiss Confederation 2012). A complete overhaul of the Act is underway (Die Bundesversammlung der Schweizerischen Eidgenossenschaft 2017) that brings it somewhat closer to its European counterpart. English translations are available for the law in force, whereas for the draft, translations from German are used in what follows.

The first question that arises is whether law and ordinance apply to the analysis of human movement data in the context of research. None of the reasons for exclusion in Article 2 apply, so the question can be reduced to whether the data to be handled is to be classified as *personal data*.

Personal and sensitive personal information

FADP and DPO make a distinction between *personal data* and *sensitive personal information* with different rules applying to the different categories. The former comprises "all information relating to an identified or identifiable person", whereas data in the second category contains more personal information, such as political views, the intimate sphere, or racial origin. The draft widens the scope of sensitive information to include, amongst others, biometric and genetic information. Additionally, there is the notion of a *personality profile* in the current version which is a "collection of data that permits an assessment of essential characteristics of the personality of a natural person". The draft instead defines the action of "profiling"

Identifiability

Therefore, to decide whether one is dealing with *personal data*, one has to decide whether or not the data subjects are identified or identifiable from the data. Most often, if the true identity of the persons of which data is available is not relevant, there are pseudonomised, which means that all information that directly points to an individual is removed and replaced with a number that carries no information. In those cases, data subjects are not *identified* but could still be *identifialbe*.

If individual trajectories are available, the question of whether they are sufficient to treat the individuals as identifiable not entirely clear based on the text of the law. The message by the federal council to the proposed new data protection act from 2017 (Der Schweizerische Bundesrat 2017) does state that positioning can be used to identify people. However, theoretical identifiability is not enough to consider the data subjects identifiable. The deciding criterion in this case is whether a party can be expected to actually make the effort to do so. Research suggests that positioning data contain a very strong signal for identifying people (De Montjoye et al. 2013), establishing the theoretical identifiability. Given the ease of re-identification, it seems cautious to assume that others would make this effort and consider the data subjects identifiable, making the data containing it *personal data*.

The following rules apply to *personal data*:

- "Personal data may only be processed for the purpose indicated at the time of collection, that is evident from the circumstances, or that is provided for by law." (Swiss Confederation 2014, Art. 4, Paragraph 3)
- "The collection of personal data and in particular the purpose of its processing must be evident to the data subject." (ibid., Art. 4, Paragraph 4)
- "If the consent of the data subject is required for the processing of personal data, such consent is valid only if given voluntarily on the provision of adequate information."(ibid., Art. 4, Paragraph 5)
- The transport of data, as well as storage and processing need to be protected from unauthorised third parties accessing, altering, deleting them. (The Federal Authorities of the Swiss Confederation 2012, Art. 9)

In addition to those rules stipulated by current legislation, the draft of the new Act on Data Protection contains multiple articles regulating the governance of data, prescribing in particular the role of a responsible person for any dataset which reflects EU terminology.

Sensitive personal information is a subset of *personal data* and thus all the above rules apply. However, in addition, also the following has to be respected when processing *sensitive personal information* and *personality profiles*:

- Wherever consent of the data subject is required, this consent must be given expressly (Swiss Confederation 2014, Art. 4).
- It may not be disclosed without justification. Possible justifications include consent or the data being made available to the public by the data subject (ibid., Art. 12).
- In general, the data subject must be informed about the data collection, the controller of the data, the purpose of processing and categories of recipients in case of a planned disclosure, even if it is acquired through third parties (ibid., Art. 14).

Exceptions for research

Breaches of privacy and the handling of *sensitive personal information* in the context of research is explicitly mentioned and can be allowed, if the person processing the data "processes personal data for purposes not relating to a specific person, in particular for the purposes of research, [...] and publishes the results in such a manner that the data subjects may not be identified" (Swiss Confederation 2014).

In summary, under Swiss law the best way to use people's mobility data is to obtain qualified express consent by the data subjects to collect and process data. The purpose of the research should be stated broadly enough to cover the intended analyses. Additionally, the data needs to be adequately protected from unauthorised third parties.

However, that all these requirements can be rendered unnecessary if the data can be collected in a way that avoids it being defined as *personal data* in the first place. The requirements are that the data subjects are neither identified nor identifiable. One way to achieve this would be to aggregate data to a level at which individuals are no longer identifiable or to use a Eulerian perspective which, in certain circumstances (Culnane, Rubinstein and Teague 2017), does not allow for the identification of the individual, as long as there is enough volume relative to the temporal and spacial granularity of the data.

GDPR

In May of 2018 the General Data Protection Regulation of the EU came into force. Because of its extraterritorial nature it applies to whomever collects or processes data pertaining to EU citizens, it also applies to the research conducted for this thesis, even if it is conducted in Switzerland.

The International Association of Privacy Professionals (IAPP) has published on its website (Maldoff 2016) an overview of the GDPR's regulations in the context of research. To summarise them, there are parallels between the GDPR and the Swiss regulation (particularly with the new draft of the FADP), not least as the Swiss legislator was aware of the GDPR when drafting the new act. As was the case for the Swiss legislation, the GDPR no longer applies to data that is not personal and it seems cautious to get consent from the data subjects wherever possible. Another parallel to the Swiss legislation draft is that the GDPR allows for the collection and processing of *personal data* even without consent for research purposes under certain conditions.

2.1.5 Non-human movement

Section 2.1 so far has been focussing on the analysis of human movement, reflecting the focus of this thesis. However, there are other types of moving entities that are being analysed. In this section, some of the similarities and differences between the computational analysis of human and animal movement are presented. While movement is analysed in many other domains as well, such as meteorology (Dodge, Laube and Weibel 2012), logistics (Patier and Routhier 2009) and medicine (Hoshiar et al. 2017) and the comparison of all involved methods would be potentially interesting, the animal domain offers a proximity that other domains typically does not: moving subjects whose movement is not (fully) determined by the physical environment.

At the beginning of any computational analysis of movement are the data. Prior to the advent of miniaturized and sufficiently small and powerful GNSS receivers, other technologies such as radio telemetry were and sometimes still are being used (Thomson et al. 2017; Toledo et al. 2018). Recent advances in miniaturisation however have opened the potential analyses to species as sensitive to weight as small bats weighing only about 20g (Dressler et al. 2016). Having a wider range of available technologies requires a balancing of length of tracking vs. frequency (if battery life is an issue) or size of the device vs. the potential area in which movement may be observed (if base stations are required to capture the signal). These are non-issues in e.g. cell phone based human tracking schemes, that can offer both relatively high tracking rates and a global scale thanks to humans' ability to recharge the tracking device. Of course, limitations of the technologies such as signal loss and bridging the gap between measured points and the underlying movement have to be handled in both the human (Fillekes et al. 2019) and the animal domain (Laube and Purves 2011).

Apart from the technologies with which to capture movement, researchers increasingly have to think carefully about how to store the data they gather. Data on human and animal movement has become so cheap to produce that it is collected in large quantities, introducing the need for big data technologies to be applied (Demšar, Slingsby and Weibel 2019; Miller et al. 2019). This problem is of course somewhat more frequently encountered in human studies due to the large number of individuals whose locations are passively tracked by cell phone operators.

In both domains, movement data has not only been recorded, but also simulated. In the animal domain, simulations are often very strongly rooted in geometry, relying on joint distributions of certain movement parameters such as speed or turning angle (Edelhoff, Signer and Balkenhol 2016; Gurarie et al. 2016; Soleymani et al. 2017) whereas human simulations are more prone to take additional knowledge about the process to be modelled into account, such as proximity to others Helbing et al. (2005) or daily routines (Balmer et al. 2009). The fact that different layers of semantics of human movement are more easily accessible than for animal movement also has an impact on the kinds of analyses that are performed. In both domains, researchers are (amongst other things) trying to solve the problem emerging from abundant and *thin* data in the sense of Miller et al. (2019) by adding semantics to them. In both domains this can include commonly used spaces, *habitats* and *activity spaces* (Fillekes et al. 2019; Wakefield, Phillips and Matthiopoulos 2009), interactions (Dodge, Weibel and Lautenschütz 2008; Helbing et al. 2005), or inferring description labels for parts of trajectories (Demšar et al. 2015; Stenneth et al. 2011). However, due to the expensive nature of ground truth labelling, even in the human domain, but of course even more for analyses of large scale animal migrations, studies using supervised classification are much more prevalent in the human domain, even if they do exist for animals too (Soleymani et al. 2014). Instead, for animals unsupervised methods or methods using rule based systems are more common (Demšar et al. 2015).

In order to generate more insights, both domains are using ancillary information wherever possible. This allows capturing the movement in more detail, such as in the case of accelerometer data (Hemminki, Nurmi and Tarkoma 2013; Nathan et al. 2012) or to capture the context in which the movement takes place, even though the nature of the captured context is different between the domains (Miller et al. 2019; Stenneth et al. 2011).

Lastly, and perhaps somewhat surprisingly, the topics of privacy and Eulerian data are not exclusive to the human domain. Eulerian data are used for animals as well, especially to describe space use (Smouse et al. 2010). In terms of privacy, especially endangered animals who are economically valuable can be put in direct danger of pouching if their movement patterns become known (Cooke et al. 2017).

2.2 Transportation mode detection

2.2.1 Context

Traditionally, transportation science has been a primary driver of the development of the computational analysis of human movement. Researchers have been interested in trip chains (Jiang, Ferreira and Gonzalez 2017), mode and purpose of trips (Zolliker, Rollier and Bosshard 2015) and OD-matrices (Ni, Wang and Chen 2018). Lately the results of the efforts in this domain are applied to fields ranging from predictive policing (Leuzzi, Del Signore and Ferranti 2017) to modelling vehicular emissions in a city (Nyhan et al. 2016) and monitoring health (Saeb et al. 2016). The most relevant aspect for this thesis is of course transportation mode detection or mode detection for short.

From geometry to semantics

In the frame set out by Section 2.1.2, mode detection is one example use case of enriching geometric information semantically. Specifically, the semantics of transport (with its stages trips, and tours) are inferred from the geometry along with other features that are not semantic themselves but are capable of revealing aspects of the semantics of the transportation mode.

Need from practitioners

The promise of (also) using technological data collection methods rather than sticking exclusively to the interviews, was first and foremost the reduction of missed trips and indeed, GNSS based surveys have been shown to record more trips than were obtained through interviews (Forrest and Pearson 2005), although the image is less clear than could be assumed, as Bricka et al. (2012) demonstrated.

In addition to the uncertainty as to whether or not a trip gets reported, there is also uncertainty whether or not the labelling of the stages making up those trips are correct. In general the reported results are very high (Feng and Timmermans 2016; Prelipcean, Gidófalvi and Susilo 2017; Shen and Stopher 2014; Stenneth et al. 2011). However, some of those results have to be interpreted with caution as they do not exactly solve the same problem that is tackled in this thesis. For example, Feng and Timmermans (2016) have calculated features based on moving windows around the fixes but separated training and testing data randomly at the fix level. Thus, the training and the testing data are not properly separated, as the moving windows from points of the training dataset and the test data set used to calculate the features overlap. This leads to results that probably would not generalise well to situations where a point, the mode of which has to be inferred, does not have a corresponding point in the training data with a large overlap in the moving window on which the features are calculated. Furthermore they use information about the person (vehicle ownership) that cannot be assumed to be available in purely passive tracking.

In the context of this thesis

The trend so far has been to add more information and better sensors to improve the quality of the prediction, which makes sense when using cell phones in the context of traditional surveys in addition to interviews and the direct contact to participants is necessary anyway. In this setting, all the sensors of the mobile phone are available and can be made use of. While some sensors, like accelerometers have been found to be useful in multiple studies (Hemminki, Nurmi and Tarkoma 2013; Nitsche et al. 2014), others such as information on wireless networks or GSM have not helped significantly with mode detection (Reddy et al. 2010).

This thesis, however, takes another approach. Instead of trying to improve results by adding sensors and sources of information, the idea is to only take the sources of information that can be assumed to be available for *passive tracking* and ask for the most important one – the positioning – what quality is required for modes of transportation to become inferable.

This section will first elucidate the differences between pointwise and segmentwise classification, and then present the most frequently used features, inferred modes, classifiers, and quality measures.

2.2.2 Pointwise vs. segmentwise classification

As described in Section 2.1.2, the semantics transport scientists are interested in are the stages and how they form trips and tours. The recorded geometries however do not provide this distinction and so, implicitly or explicitly the trajectories to be labelled have to be partitioned into stages.

Stages and Segments

The title of this section implies the existence of *segments*, which have not been mentioned thus far. A segment of a trajectory is a temporally contiguous part thereof and thus is a geometric concept, whereas a stage, defined in Section 2.1.2 clearly is a semantic concept.

Why, therefore, is there a distinction between stages and segments? The reason can be found in the way classification is performed and for that some clarity of what meant by mode detection is needed.

Mode detection

Given $N \in \mathbb{N}$, a trajectory $(id, t_i, x_i, y_i)_{i=1}^{N}$ and a list of transportation modes $\mathbb{M}$, mode detection finds a number of stages $p \in \mathbb{N}$, modes for those stages $m_1, \ldots, m_p \in \mathbb{M}$, timestamps marking the beginnings and ends of the stages $t_1^c, \ldots, t_p^c \in [t_1, t_N]$ such that for the tuples $(t_1^c, m_1), \ldots, (t_p^c, m_p)$ it holds that $t_p^c = t_N$ and if $p > 1 : \forall j \in \{2, \ldots, p\} : t_j^c > t_{j-1}^c \wedge m_j \neq m_{j-1}$.

In words this means that mode detection partitions the time covered by the trajectory (which may only be part of the time where the actual movement took place) into parts with different transportation modes, i.e. stages. If $p = 1$ for example, the mode detection has detected only a single stage.

A visual representation of the task is presented in Figure 2.4. The semantics (top row) is the ground truth partition of the trip into stages that is to be recovered by mode detection. The second line shows the fixes, i.e. the raw data that are collected. They may of course contain many more sensor readings (features) than just the ones depicted, but the ones in the image show the bare minimum of what is always available. The last line finally shows the result of the mode detection, consisting of deciding where to insert breaks between the inferred stages and, inferring the labels of those inferred stages.

In this process there are several things that can go wrong, as also indicated in Figure 2.4. First, breaks can be inserted where in the ground truth there are none (the red inferred stage is superfluous), the exact position of the inferred stage breaks can be shifted with respect to ground truth (the one between the blue and the turquoise stage) or a stage break can be missed entirely (the one between the turquoise and the purple stage). It is important to note that the inferred information always starts at the first and ends at the last fix. Any movement that happens outside this temporal window is not recorded and can therefore not be discussed based on the available data. A final point to make is that the number of fixes can be(and typically is) significantly larger than the number of stages. The reduced number shown simply makes the representation more readable.

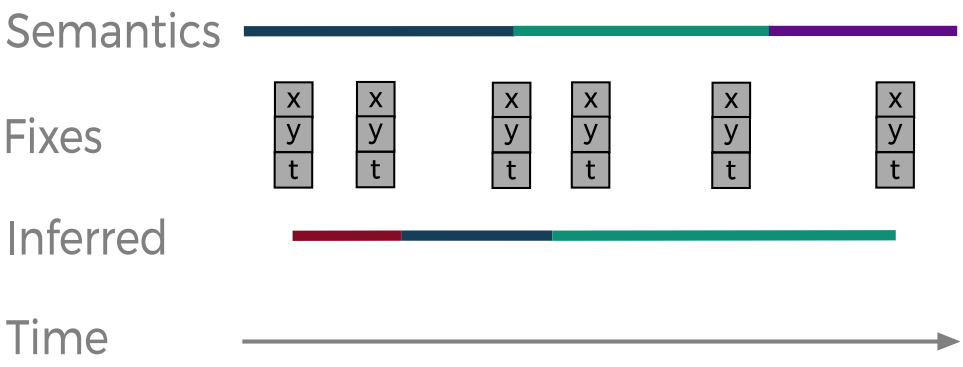


Figure 2.4:
Visualisation of the transportation mode detection process. The true modes of traffic to be inferred are in the first row, the measured fixes on the second and the inferred modes on the third. The inferred labels need not coincide semantically or temporally with the ground truth and their temporal extent is that of the data, not that of the ground truth.

The final step of inferring the labels is in most cases tackled as a supervised learning problem, where mappings between features of parts of the trajectory and labels are learned. Consecutive parts bearing the same label are then merged into a single stage to obtain the final result.

Pointwise

The pointwise mode detection paradigm performs classification on the granularity of the fixes. This restricts the possible mode change points, most often to the timestamps of the points themselves (Reddy et al. 2008) or – less commonly – to the midpoints between the timestamps of fixes (Bolbol et al. 2012).

This assumes that the modes of transportation are separable in a space of features which can either be measured directly for every fix or calculated and attributed to each fix. One example of such a feature, which is commonly available through GNSS devices is speed, which is used often and has been found to be very informative when differentiating transportation modes with some studies relying almost entirely on this feature (Schuessler and Axhausen 2009).

GNSS sensors, which constitute the most common data source for this task are typically used at a sampling rate that is significantly higher than the rate at which modes of transportation change. To find expressive features, it can therefore be useful to include information about temporally 'close' features, as the mode is likely to be the same, allowing for a reduction in variance. The two ways most commonly used ways to add this information is either using moving windows (Stenneth et al. 2011) or segmenting the trajectory (Chen et al. 2010).

If moving windows are used, some of the calculated features of a fix are based on the features of neighbouring fixes. The broader the windows, the more the features will look similar on two neighbouring fixes. While it is possible to use multiple windows simultaneously, for example one to capture short-term patterns and one to capture more long term aspects of the movement, or one looking backward and one looking forward, usually only one single moving window is used.

Segmentwise

Segmentation on the other hand, tries to use domain knowledge to find contiguous parts of the (geometric) trajectory that are then assumed to have the same label. Features are then calculated based on all the fixes that fall into a single segment and a single label for the entire segment is inferred by the mode detection procedure. It is important to note that segmentation happens *outside* the learning process and thus is performed using rules set by the analyst. The mode detection procedure only sees information at that granularity and is not capable to split up the segments afterwards. For example, Sauerländer-Biebl et al. (2017) use thresholds on speed as the boundaries of segmentation, arguing that between any two modes the speed must drop to (almost) zero.

Def. Segmentation

More formally, given a trajectory $(id, t_i, x_i, y_i)_{i=1}^{N}$ for some $N \in \mathbb{N}$ segmentation finds a number of segments $p \in \mathbb{N}$ and time stamps $t_1^c, \ldots, t_p^c \in [t_1, t_N]$ such that $t_1^p = 1, t_p^c = t_N$ and if $p > 1$: $\forall j \in \{2, \ldots, p\} : t_j^c > t_{j-1}^c$.

If segmentation is used, all identified stages start and end where a segment starts or ends, but not necessarily vice versa. Segmentation is thus a partition of trajectories on the geometric level and restricts the boundaries of the inferable stages of the semantic level. Thus, it is a further restriction of the positions of the t_j^c in the definition of mode detection provided previously. From a process point of view, segmentation is usually the result of a domain knowledge informed, rule based procedure, mostly containing at least some element of identifying parts of the trajectory with (near) zero speed (Chen et al. 2010; Gong et al. 2012; Huss et al. 2014; Pereira et al. 2013; Sauerländer-Biebl et al. 2017; Zhang et al. 2012). Some authors have used a segmentation that was given by the stages (Bohte and Maat 2009; Bolbol et al. 2012), but the assumption of course limits the application to situations where those stages are known at the time of inference, which is not the case in passively sensed mobile phone data. The inference of the modes of those segments themselves on the other hand is mostly the result of some optimisation (Shen and Stopher 2014). The difference between point-based and segmentation based mode detection is visualised in Figure 2.5.

One method that falls somewhere between pointwise and segmentwise classification is that of conditional random fields (CRF) on pointwise features. While the classification is clearly pointwise, the fact that this method typically learns that the same label repeats itself leads to longer sequences of identical labels. Overall, CRF's have been found to perform worse than two-stage approaches (Zheng et al. 2008). Other alternatives, such as recurrent neural networks (Lin et al. 2017), sequence to sequence models (Sutskever, Vinyals and Le 2014), or attention based classifiers (Vaswani et al. 2017) are conceptually very well suited for the task, but have so far not been the focus of research in mode detection.

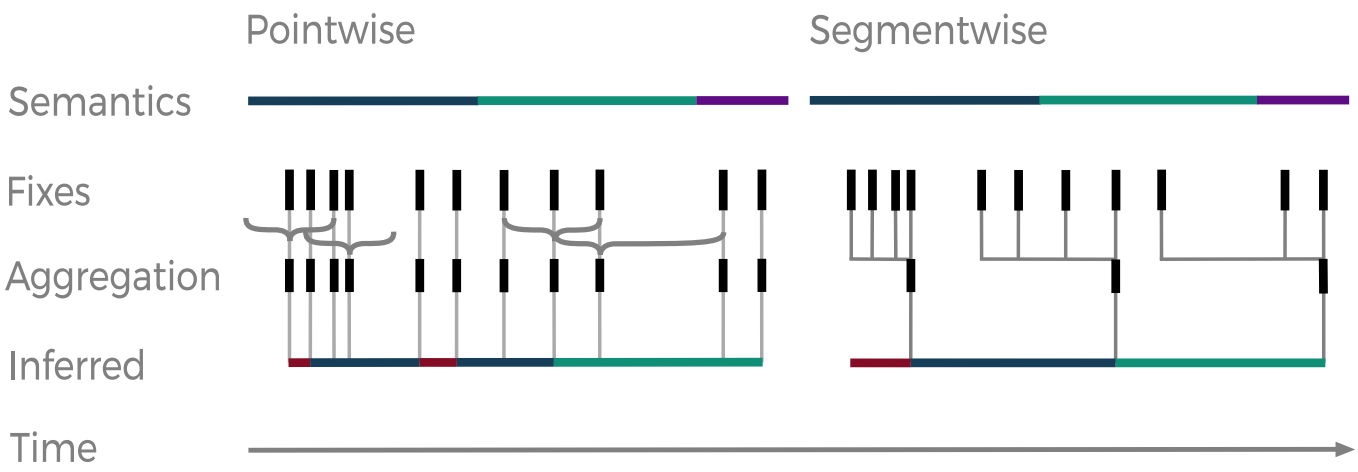


Figure 2.5: Difference between pointwise and segmentwise detection of modes of transport. In pointwise classification, the fixes delineate the possible breaks of the inferred stages. The original fixes may be enhanced by using summary statistics from 'nearby' features. Nearby can mean all fixes falling in a fixed time window (left two braces), or the closest n fixes for some – typically small – n (right two braces). Every aggregated feature vector is then used to classify the time between two consecutive fixes. The overlap of aggregation windows can lead to slowly changing aggregated feature vectors. In segmentwise classification, rigid segments are defined first, and, for those segments, entirely new feature vectors are calculated based on all fixes of a segment. One single label is then inferred for the entire segment. As shown, in pointwise classification, the features of every fix may end up contributing to multiple aggregated feature vectors whereas in segmentwise classification, every fix contributes to at most one aggregated feature vector.

Why segment

It is reasonable to ask why a researcher would desire to restrict the possible transportation mode change points using a segmentation; after all they run the risk of masking a true change point that can never be recovered by the mode inference. There are two arguments in favour of doing so:

First, they are based on domain knowledge and if they are chosen to actually contain only one transportation mode, features calculated on the segment can be more expressive for that mode. For example, if a segment captures a bus ride over many stops, features that express the typical stop-and-go behaviour of a bus can be helpful in mode detection.

Second, pointwise mode inference is noisy. While it is true that features of nearby points can be assumed to be close in feature space, particularly if the features include some that were calculated over large moving windows, there is still a tendency of having some stray points with wrong labels in a trip. This of course inflates the number of inferred stages, witch is definitely not desirable.

This second point however, can be mitigated somewhat by introducing a smoothing step after the initial classification. Most studies seem to be using smoothing schemes based on either hidden Markov models (Nitsche et al. 2014, 2012; Reddy et al. 2010; Shah et al. 2014) or a simple majority vote in a moving window (Prelipcean, Gidofalvi and Susilo 2016).

2.2.3 Features for classification

A very wide range of features has been used to infer the mode of transportation. In this section they will be presented in groups representing their origin, as many features rely on e.g. a specific sensor to be available.

Features from localising sources

The localising sources that are most commonly used for transportation mode detection are clearly GNSS sensors, and in a few studies, passive mobile phone tracking.

As discussed in Section 2.1.3, GNSS devices provide not only timestamped positional estimates, but also a host of other measurements, most notably velocity, estimates around the (un-)certainty of the positioning, and the number of satellites that have contributed to the estimate and altitude.

All of them have been used directly in prediction. Speed can intuitively contribute a great deal to separating transportation mode and is the feature which has clearly been used most often. Speed-based measures include averages (Bohte and Maat 2009; Gonzalez et al. 2008; Schuessler and Axhausen 2009; Stenneth et al. 2011; Stopher, FitzGerald and Zhang 2008; Zhang et al. 2012; Zheng et al. 2008), extremes (Bohte and Maat 2009; Gonzalez et al. 2008; Nitsche et al. 2014, 2012), quantiles (Gong et al. 2012; Huss et al. 2014; Nitsche et al. 2014, 2012), acceleration (Bolbol et al. 2012; Huss et al. 2014; Schuessler and Axhausen 2009; Shafique and Hato 2015; Zhang et al. 2012; Zheng et al. 2008) and variability (Zheng et al. 2008)

Other features based on GNSS that are used in transportation mode detection are precision or noise (Ellis et al. 2014; Gong et al. 2012; Stenneth et al. 2011; Xiao, Juan and Zhang 2015) and number of satellites (Gong et al. 2012; Jahangiri and Rakha 2015).

Apart from direct sensor readings, a wide range of derived features based on the timestamps are also being employed, including displacement (Feng and Timmermans 2016), turning angle (Nitsche et al. 2012) and self-intersection (Prelipcean, Gidofalvi and Susilo 2016).

It is perhaps somewhat astonishing that staples of the computational analysis of animal movement, such as tortuosity (Gurarie et al. 2016) are not widely featured in research about mode detection. Drawing upon the very geometrically focused research on animal movement could perhaps turn out to be beneficial even in the context of traffic.

In order for GNSS sensors to produce fixes, they need to receive the signal from the satellites. If the view of the satellites is obstructed, this can lead to signal loss for parts of the movement. This can be a frequent occurrence, particularly in transportation modes such as underground or trains (Widhalm et al. 2012). However, this needs not only be a problem and can actually be a useful signal of its own as long as signal loss occurs in clear patterns for certain modes of transportation. Thus far, this has not received extensive coverage in the literature.

Features from non-localising sensors

Although this thesis does not rely on non-localising sensors, omitting them completely in describing the literature on mode detection would not do justice to their importance.

Non-localising sensors reverse some of the abstraction of human movement as translation of a single, nondescript point in space and bring the analysis somewhat closer to the micro-scale. Because all transportation modes have phases of slow movement it may be difficult to distinguish walking, slow cycling, and slow driving based on geometry alone. In such cases, adding information beyond the geometry of a trajectory can help.

Most important of those sensors is the accelerometer that provides the values of acceleration in and/or rotation around the three spatial axis. This is particularly useful to distinguish passive, sedentary modes of transportation from active ones such as walking where acceleration changes more and more rapidly. Adding accelerometer-based features, where available, has thus benefited the quality of mode detection, where available (Hemminki, Nurmi and Tarkoma 2013).

Features outside the human in motion

Movement does not happen in isolation, but (mostly) in relation to a geographic environment, often in the form of a built infrastructure. Cars mostly need roads and trains need tracks, so knowing the relationship between the trajectory and the relevant infrastructure can contribute to the quality of the inference.

In most cases where static information that holds for all movement is available, at least the distance of a fix to the closest stop on the various modes of public transportation, or even the closest line of different public transportation systems are added to the features (Chen et al. 2010; Gong et al. 2012; Moiseeva and Timmermans 2010; Semanjski et al. 2017; Stenneth et al. 2011; Zhu et al. 2016). Nowadays, open data such as OpenStreetMap [7] have such a good coverage in many countries, particularly in urban areas, as Ferster et al. (2018) showed for bicycle infrastructure, those distances are available in most study areas and even the roads can be split up into different classes (Semanjski et al. 2017).

While already knowing the distance to the infrastructure can be a great boon to the prediction (Chen et al. 2010), knowing the distance to the actual vehicles can be even more helpful, as was impressively demonstrated by Stenneth et al. (2011). However, this kind of information is not always easily accessible and relies mostly on open government data such as the one available in Switzerland [8].

Several studies have shown that not only the stops and lines of public transportation are useful, but also other pieces of infrastructure, such as observable Wi-Fi stations (Reddy et al. 2010).

[7] www.openstreetmap.org

[8] https://opentransportdata.swiss

While knowing the infrastructure – transportation and otherwise – is helpful in adding the semantic level to the geometry of the trajectory, already knowing parts of that semantic level can also be helpful. In some studies, information on the users themselves are available e.g. on the ownership of bicycles (Feng and Timmermans 2016; Stopher, FitzGerald and Zhang 2008). This kind of information has a direct effect on the propensity of said users using those modes of transportation and can help to improve inference. However, it can not be assumed that this information would be available on a larger scale for use with passively sensed mobile phone data.

2.2.4 Modes of transportation

Different modes in different regions

When it comes to selecting $\mathbb{M}$, the set of modes from which the inferred labels have to be selected, there is unfortunately no consensus as to what should be included. This is not surprising, given that different compositions of transport modes are relevant in different regions. This makes a fair comparison of the results much more difficult: If in a region there are only pedestrians and private cars it is much easier to distinguish all relevant modes of transportation than in a city where there are also significant numbers of bikes, buses, and trams, which in cities may all have a speed profile similar to that of the car. Furthermore, it is not entirely clear from the literature how easily transferable classifications learned on one dataset are to another.

This means that, in every region and for every application where a mode detector is to be applied, there is a need to collect high quality labels to either ensure that a pre-trained model also works for the area in question or to train a classifier on the data itself.

Despite all this, there are some modes that are fairly common in most articles on the topic: walking, driving (car), cycling and riding a bus can be found in most of them. Trains are already less frequent and others, such as trams, motorbikes, general motorised travel, running, or e-bikes all constitute a rarity.

There is also some taxonomical uncertainty about the beginnings and ends of stages. In particular, it is not entirely clear how the time between two stages should be treated. While there has been research into the exact nature of how such transfers look (Das, Ronald and Winter 2014), in most studies there are no explicit transfer stages.

2.2.5 Classifiers

A broad variety of different classifiers are used to detect modes of transportation. From the domain knowledge informed rule-based approach (Bohte and Maat 2009; Chen et al. 2010; Das and Winter 2016, 2018; Gong et al. 2012; Sauerländer-Biebl et al. 2017; Schuessler and Axhausen 2009; Stopher, FitzGerald and Zhang 2008) to traditional methods such as k-nearest neighbours (Jahangiri and Rakha 2015; Reddy et al. 2010), support vector classification (Bolbol et al. 2012; Pereira et al. 2013), decision trees (Reddy et al. 2010), random forests (Ellis et al. 2014; Mäenpää, Lobov and Martinez Lastra 2017) everything can be found. Surprisingly the number of deep learning based algorithms has so far been rather limited in relation to the impact it had on traffic flow prediction (cf. Section 2.3).

In addition to those direct methods that take a feature vector (of a point or a segment) and attribute a label to it, there are some methods that try to incorporate the sequential nature of the trip to be classified. This happens either by applying some hidden Markov model after an initial classification (Bantis and Haworth 2017; Nitsche et al. 2014; Reddy et al. 2008) or by using methods that consider the sequential nature of the underlying movement directly, for example by applying CRF's to the feature vectors (Zheng et al. 2008). Recurrent neural networks, that also are adept at labelling sequences (Huang, Xu and Yu 2015) have not yet featured prominently in the literature.

Within the work in this area there is no one classifier that seems to dominate the others, as conflicting results about the relative performances can be found. Consequently, it may be advisable to test several classifiers for any given research or application problem.

2.2.6 Quality measures and comparability

Different measures for different needs

In many cases, only precision, recall, accuracy and/or the F1 score are being reported (often by transport mode). However, especially in the case of unbalanced data (i.e. vastly different frequencies for the different modes), a high accuracy does not necessarily mean good classification. Therefore, some authors have also used either Cohen's Kappa (Bolbol et al. 2012; Huss et al. 2014) or the Chi-Squared (Bantis and Haworth 2017). Other metrics, such as the ones proposed by Prelipcean, Gidofalvi and Susilo (2016) have yet to be widely adopted.

While those quality measures are common in machine learning and helpful in assessing the qualities of a given classification scheme, they are not necessarily the measures that transportation researchers have been using. In the study by Houston, Luong and Boarnet (2014) for example, the primary interest lies in the number of trips and total time per transportation mode.

In small studies that only assess the performance of a classification scheme the traditional measures used in machine learning are useful to capture the behaviour of those schemes. In view of the application of those methods to questions of transportation research it may be useful to also consider what the important quality measures are in that domain.

Incomparable research

A problem plaguing the field of transportation mode detection is the lack of comparability. While some datasets such as Geolife (Zheng, Xie and Ma 2010) have been used quite frequently, for the most part researchers have used their own data. This can be highly rational, because in different cities different mode splits can lead to different performances and some important modes for the study area of the researchers may not be available on public datasets.

In addition to the geographical problems inherent in the nature of the research domain, there is also the fact that even slightly different problem statements can mean that the same key figures such as a confusion matrix on the same modes of transport can mean very different things. Whether, for example, the confusion matrix is by time or by fix, by segment or by trip has implications regarding what values can be seen as acceptable. Furthermore, there are differences in the history that are assumed to be known. For location based services, only the history prior to the point to be classified can be assumed to be known. In transportation science on the other hand, the whole trajectory is typically known, and information from the "future" can also be used in classification (Prelipcean, Gidófalvi and Susilo 2017).

While the problem of different sets of transportation modes in different regions will probably remain and force researchers to always obtain their own (high quality) data labels, the problem of agreeing on a (small) set of precise problem definitions in the field of mode detection would go a long way towards making results at least somewhat more comparable.

2.3 Traffic flow prediction

One of the fields that may be heavily affected by a shift to passive tracking is short term traffic prediction. This is a problem that has been studied for decades (Lighthill and Whitham 1955) and is concerned with forecasting aspects of traffic in the short term. This can then either be used to plan ideal routes depending on time-varying traffic conditions (Kok, Hans and Schutten 2012) or control traffic by setting speed limits or switching red lights (Papageorgiou et al. 2003).

Fundamental relationships

Different aspects of traffic are relevant, depending on the application. In the context of prediction, most often speed, travel time, or flow (volume) have been the focus of research (Vlahogianni, Karlaftis and Golias 2014). These different variables are interconnected with each other in the so called fundamental diagram of traffic flow, which describes the relationship between flow, density and speed on a street. The relationship is given by $q = \rho \cdot v$, where q denotes the traffic flow (vehicles by time interval), ρ denotes the density on the street (vehicles by distance) and v denotes the speed of the vehicles (distance by time). A depiction of the possible realisations of a steady state are shown in Figure 2.6. If the density is zero, then so is the flow. The first few vehicles are not restricted by others and thus their speed is limited only by factors such as the speed limit for the street. Increasing density will lead to lower headway between vehicles and therefore reduce the speed at which they will drive. Eventually the decrease in flow due to reduced speeds will outweigh the increase in flow due to increased density. At this point the street has reached its maximum capacity. Any increased density beyond this point will lead to a reduction in flow up to the point of a jam.

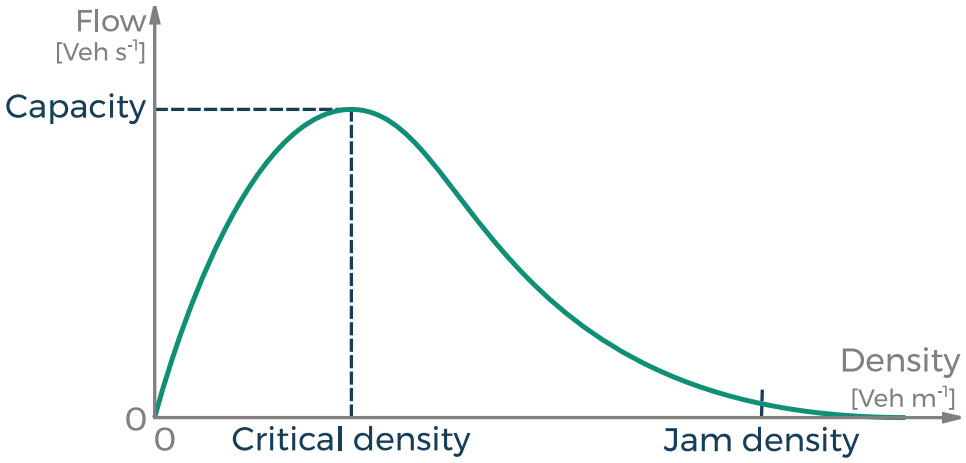


Figure 2.6:
Fundamental diagram of traffic flow. There is a critical density allowing for a maximal flow. After this point the reduction of speed incurred by more vehicles outweighs the increased density, reducing overall flow. Figure adapted from Ferrara, Sacone and Siri (2018)

Predictions of the components of the fundamental diagram can help with optimal routing and are therefore of importance. There are different approaches to that kind of prediction. The first is bottom up and models the individual participants of traffic. One example of this is agent based modelling (Balmer et al. 2009). Other approaches very deeply rooted in transportation science that are used to predict the short term future of traffic spend great care on the dynamics and interactions of individual vehicles as well as with other traffic related entities such as traffic lights (Nantes et al. 2016). Those approaches can be incredibly accurate because they have access to very fine grained and accurate data, both temporally and spatially. Such data are not typically available for long durations across large areas – even with passive tracking – and therefore these kinds of approaches are not that wide spread. Furthermore, their predictions are in the very short term (Knospe et al. 2000) and therefore are only of limited use when predicting in the time windows in the orders of magnitude people use to plan their journeys (minutes to hours).

An alternative, more scalable and common method, however is to use historic data on traffic conditions and using them to predict future conditions. For this method, researchers usually rely on a network of loop detectors covering the region in question. As noted in Section 2.1.3, even many western regions including the one around Zurich do not offer such data openly in high temporal and spatial resolution and this is precisely where the passive tracking of large sections of the population could turn out to be helpful.

Once semantics such as mode of transportation have been added to the (highly sensitive) information at the granularity of the individual, the data can be Eulerised and provided as unproblematic data at relatively high spatial and temporal granularities, for all relevant modes of transportation and the changes between them. This would provide a coherent view on multimodal travel that has not yet been possible.

It is true that companies like Google can and do provide services based on the large scale tracking of users, but they do not readily share data and the depth of data they share is limited. Therefore there is room for another potential source for this kind of information.

2.3.1 Traditional statistical approaches

A common way of looking at traffic flow prediction, especially if the data comes from a fixed set of sensors, is as one of predicting a time series, which happens on a larger and coarser scale than the bottom up methods described above. The data consist of or are aggregated into chunks of between one (Sheu, Lan and Huang 2009) and sixty minutes (Zhong, Sharma and Lingras 2005), which prohibits a fine grained analysis based on advanced transportation science. Moreover, as the data are now clearly Eulerian, interactions between vehicles can no longer be modelled individually. However, broader geographical scope and further prediction horizons can be achieved this way.

Time Series Analysis

A traditional way to approach the time series framing of traffic flow prediction is by using variants of ARIMA models, such as simple ARIMA (Ahmed and Cook 1979), seasonal extensions thereof known as SARIMA (Szeto et al. 2009), some multivariate variants like VARIMA and STARIMA (Kamarianakis and Prastacos 2003) or extensions explicitly modelling the 'downstream' and 'upstream' stations such as in ARIMAX models (Williams and Hoel 2003). These methods, broadly speaking, predict the next value of a sequence as a linear combination of previous values, while modelling the error term as a moving average of preceding error terms, thus allowing for temporally correlated errors (Shumway and Stoffer 2010). While these models have been popular in the past, in more current literature they feature mostly as comparisons or baseline against which newer models are compared.

Another important family for traffic prediction is filtering, using Kalman filters (Okutani and Stephanedes 1984; Whittaker, Garside and Lindveld 1997) or particle filters (Chen, Rakha and Sadek 2011).

2.3.2 Multivariate vs. univariate

Univariate predictions have been very common for many years (Vlahogianni, Karlaftis and Golias 2014; Vlahogianni, Golias and Karlaftis 2004). This means that the volumes or speeds at a single station have been predicted. Furthermore when multivariate extensions to ARIMA models were tested against univariate ones they were found to be inferior (Kamarianakis and Prastacos 2003). This is somewhat surprising as intuitively one would expect the flows of other station to be able to help predict the flows.

Most of the multivariate predictions, however, concerned themselves only with very small geographic regions, such as stretches of one highway or arterial (Wu and Tan 2016). While this may be quite beneficial for very short-term predictions, intuitively one would not expect this to be helpful for somewhat longer prediction horizons, that exceed the time it takes to traverse the stretch used. Other approaches disregard the topology of the traffic network and treat 'neighbourhood' relationships in the data matrix as semantically meaningful (Liu et al. 2017). Although this reduction of space to the position in a data matrix apparently improves the error rate, it does come with the downside of being oblivious to the differences between, for example, a sensor on a highway ramp and a sensor closeby on the opposite side of the road which can be problematic, at the very least conceptually.

Few global models

Truly global models that incorporate some information on the global state (for predictions further out) as well as local information (for short-term predictions) are extremely rare. One reason for this is that they would need to be able to adapt to changes in the topology of the network and handle large input matrices, which un turn would lead to very high numbers of trainable parameters, thus slowing down the training process. One promising way of dealing with such a large number of inputs is using graph based methods that look at topological neighbourhoods. This can dramatically reduce the number of parameters, even if information on the system as a whole may get lost (Shahsavari and Abbeel 2015; Yu et al. 2017), as there are no more absolute positions.

2.3.3 Covariates

Most methods used in the literature, in particular the ones based on time-series methods, only use past traffic flows to predict future traffic flows. However, it seams reasonable that other information can be useful when predicting the flows.

Sometimes, models are fitted on weekdays only, excluding any public holidays (Polson and Sokolov 2017; Stathopoulos and Karlaftis 2003). In practice, this would mean that for weekends and/or public holidays separate models would have to be fitted. As those models for special days would then need to be fitted on significantly fewer data, the impact of the limited size of their training set on their quality would have to be carefully analysed.

As with any prediction, knowing about the context can be useful. Instead of ignoring special days, it would be possible to present this information as a boolean covariate that could be used by a predictor. In a similar vein, other context variables for traffic, such as weather related information (Elhenawy and Rakha 2016) could be used but this is rarely done.

2.3.4 Deep learning based methods

In recent years, many approaches to improve on existing models for traffic flow prediction are based on deep learning. The book by Goodfellow, Bengio and Courville (2016) provides an excellent introduction into the topic. For reasons of completeness a brief summary of the most important concepts are presented here to at least position deep learning within artificial intelligence and introduce some basic terms.

As illustrated in Figure 2.7, *classic machine learning* is generally concerned with learning an input - output relationship. Because sometimes raw data are not easy for the machine learning algorithms to deal with, features, i.e. variables derived in a deterministic manner from the raw data can be used instead as input to the mapping. In *representation learning*, the features are not chosen deterministically by the analyst, but instead learned to improve the overall input - output mapping. In *deep learning* finally, features are extracted not only from the original input data but also from other features, creating higher order features.

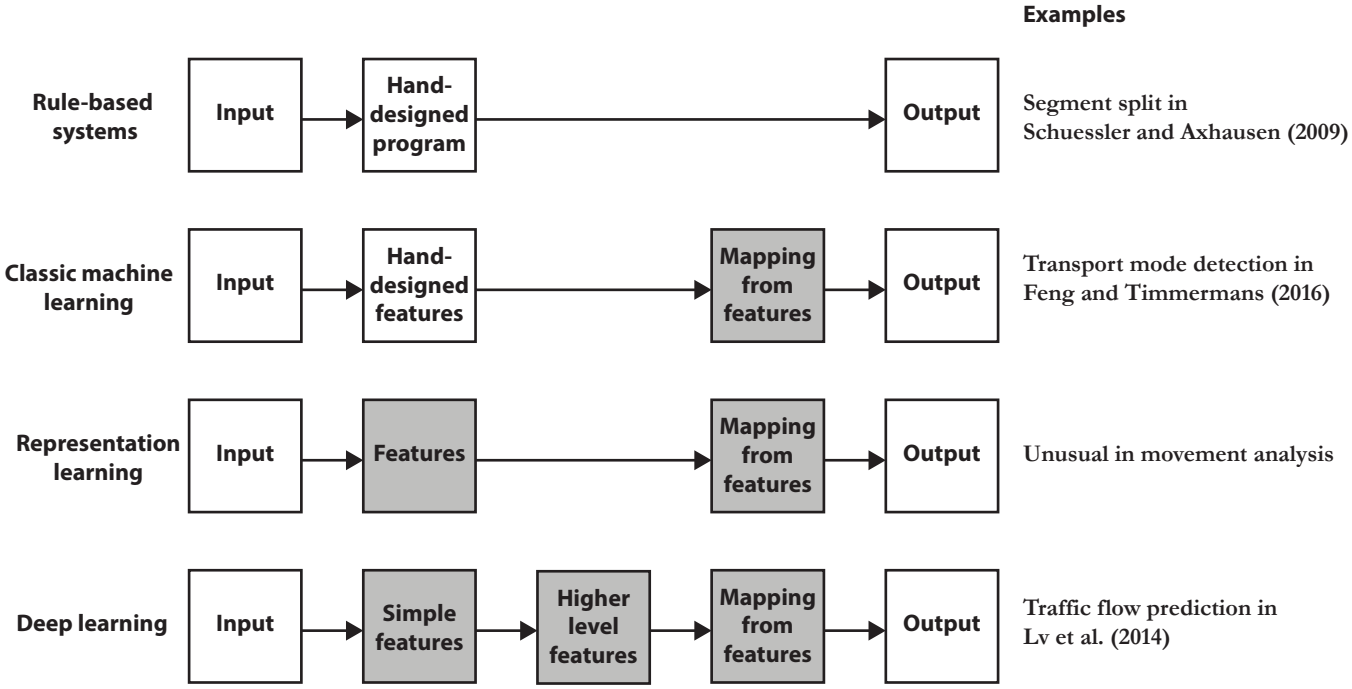


Figure 2.7: Different approaches to inferring information on data. Grey boxes represent stages in the process where learning takes place, i.e. where only the structure of the mapping from one stage to the next is determined and not the mapping itself, which has to be learned from the data. Adapted from Goodfellow, Bengio and Courville (2016).

Deep artificial neural networks are one way to perform deep learning. The fundamental example here is the *Multilayer Perceptron* (Rumelhart, Hinton and Williams 1985) illustrated in Figure 2.8. Every circle that is neither in the leftmost nor rightmost layer represents the equation $x_i^{(l)} := f\left(\beta_{0,i}^{(l)} \cdot \sum_k x_k^{(l-1)} w_{ki}^{(l)}\right)$ where the index $l \in \mathbb{N} \setminus \{1\}$ denotes the layer, k runs over the indices of the values in the previous layer, $\beta_{0,i}^{(l)} \in \mathbb{R}$ denotes an offset (to be learned) , $w_{ki}^{(l)} \in \mathbb{R}$ is a weight (also to be learned) and $f : \mathbb{R} \to \mathbb{R}$ is a nonlinear function, often the sigmoid, $f(x) := \frac{1}{1+e^{-x}}$, tanh or ReLU, $f(x) := \begin{cases} 0 & \text{for} x < 0 \\ x & \text{for} x \geq 0 \end{cases}$.

The leftmost layer ($l = 1$) represents the input and hence no calculations are needed. The rightmost layer on the other hand depends on the nature of the problem to be solved. In case of regression, there is often only a single node present and instead of a nonlinear function the identity, $f(x) = x$, is chosen. In case of a classification problem on the other hand, the number of nodes in the last layer corresponds to the number of different classes between which the mapping should be able to distinguish and it is ensured that the values sum up to one, which allows for an interpretation of the output as probabilities of belonging to one of the classes.

Note the correspondence between the bottom row of Figure 2.7 and Figure 2.8. Input and Output are labelled explicitly. The simple features are found in the first layer of the MLP, the higher level features are the values in the higher order layers and the mapping from features to output happens in the last layer before the output.

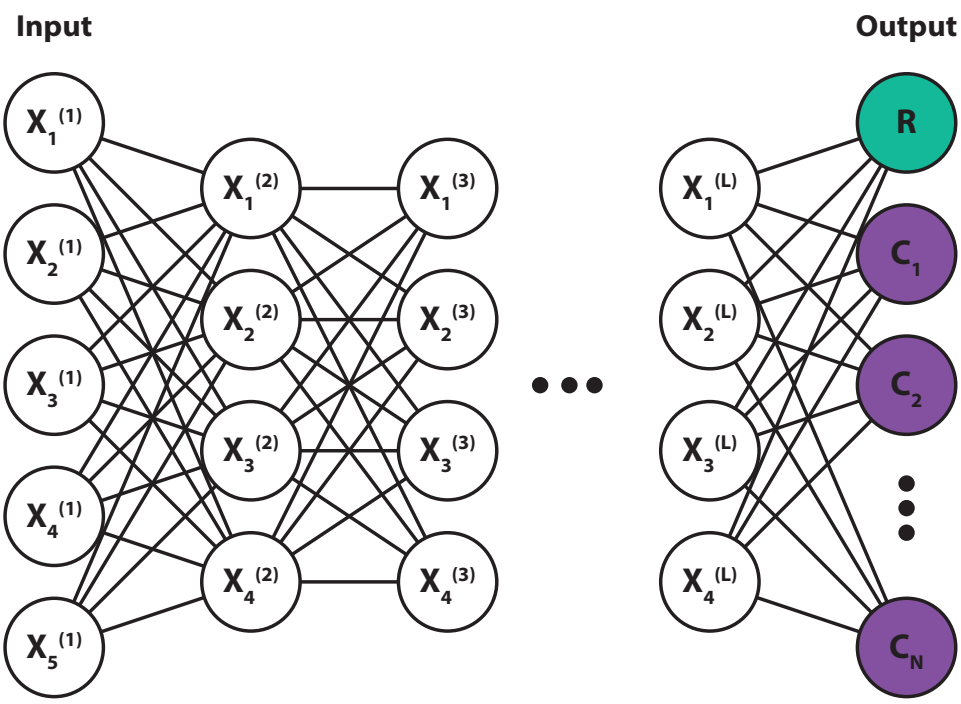


Figure 2.8:
An example of a multilayer perceptron with 5 input variables, $L \in \mathbb{N}$ hidden layers with 4 nodes each. In a regression problem, the output would be a single node, illustrated by the green R, whereas in a classification problem with $N \in \mathbb{N}$ classes there would be N nodes in the final layer (illustrated in purple), with values in $(0, 1)$, adding to unity.

The actual *learning* in the name deep learning then refers to the process by which the parameters $\left(\beta_{0,i}^{(l)}\right)_{i,l}$ and $\left(w_{ki}^{(l)}\right)_{i,k,l}$ are found that lead to a mapping that reduces the difference between the inferred output and the true output captured in the *error function*. Often, this is achieved by a process called *backpropagation*: Since all the building blocks of the multilayer perceptron are differentiable, the chain rule allows the derivative of the error function with respect to the parameters to be found by passing the partial derivatives backwards from the last layer to the first. This allows adjusting the parameters in a direction in which the error function is assumed to decrease.

One problem with multilayer perceptrons is that the number of parameters to be learned grows very quickly with increasing depth and increasing numbers of nodes within each layer. Different architectures thus have been devised to reduce the complexity of the models while retaining their expressiveness.

Feed-Forward networks

Unlike the field of transportation mode detection the field of traffic flow prediction has been affected strongly by the surge in popularity of artificial neural networks. The more common feed-forward and recurrent architectures have been tried in variations such as single layer networks (Smith and Demetsky 1994), multilayer perceptrons (Lv et al. 2014) and stacked autoencoders, in which a multilayer perceptron is pre-trained layer by layer, to allow for better predictions, the fuzzy-neural approach, in which different parts of the network learn to predict for different states and attribution to the stages is controlled by a gate at the beginning (Yin et al. 2002), and the spectral basis network, in which the input is transformed into a higher dimensional space in which there is the promise of increased separability (Park, Rilett and Han 1999).

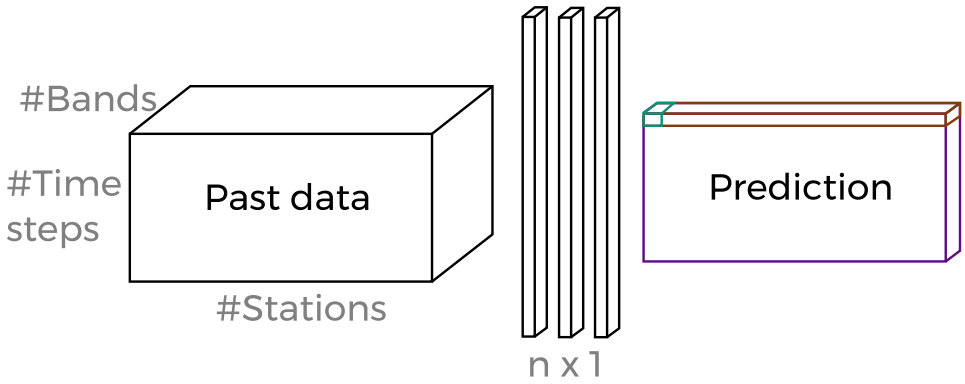


Figure 2.9:
The feed-forward architecture used in this study. The absence of any imposed structure in the data results in a multilayer perceptron. The univariate one time step prediction only predicts the turquoise cube, whereas the multivariate one time step prediction predicts the brown cuboid. The multi time-step multivariate prediction has to predict the entire purple cuboid.

These feed-forward networks all take a given part of the past and usually predict a single value for one or more stations at a single point in the future, as illustrated in Figure 2.9. This limits the information that can be used for prediction to the given part of the past. Anything that happened before that time is ignored in the prediction. Addressing this problem by making more of the history visible to the networks ever larger will strongly increase the number of parameters. This is due to the fact that all these architectures share the limitation that they do not encode some structure in the data using weight sharing. Because every station is treated individually and no convolutions are applied, the number of parameters quickly grows relatively large and becomes concentrated in the first layer, especially if there are many bands or if all time steps from the preceding week to the current point in time are provided as input.

One way of addressing this is by selecting only a selection of time steps, as temporal contiguity is not a requirement. For example, providing only the traffic conditions one day and one week before the time of the prediction has been done, reducing the necessary time slices from 673 to 3 (Wu and Tan 2016). However, it can be argued that this only captures the hebdomadal and diurnal seasonal effects and will not be able to handle idiosyncrasies of a given situation.

LSTM's

A conceptually different approach is that of recurrent neural networks, as described in Goodfellow, Bengio and Courville (2016), in particularly ones based on LSTM (long short-term memory) (Hochreiter and Schmidhuber 1997) or GRU (gated recurrent unit) cells (Cho et al. 2014). In these architectures, the input sequence is processed one step at a time, allowing certain memory cells to store information for as long as it is needed. This helps to address the problem of limited history, as information is only 'forgotten' once it is deemed irrelevant. Instead of feeding the network the entire input of past data depicted in Figure 2.9 at once, every time step (i.e. horizontal slice) is vectorised and input at the bottom of an LSTM based network, as depicted in Figure 2.10. Instead of simple neurons, every hidden layer of such a network uses so called LSTM cells. Following the notation used by (Zaremba, Sutskever and Vinyals 2014), an LSTM cell can be described as follows.

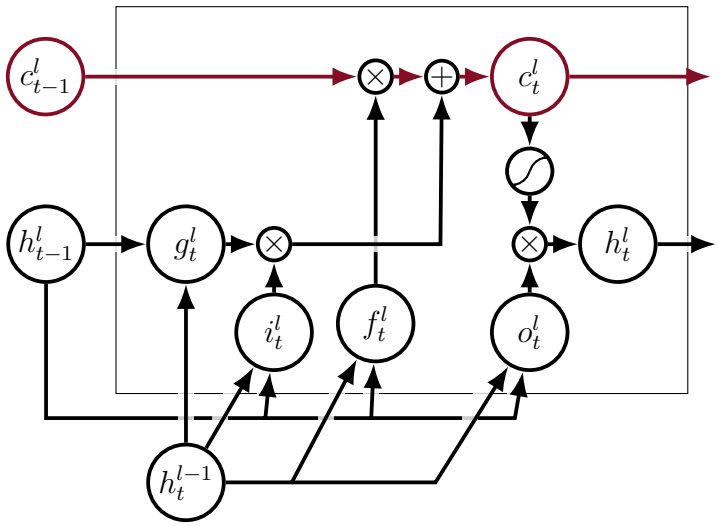


Figure 2.10:
Visualisation of the calculations in an LSTM cell. The red path is the way through which information can be retained over many time steps. As long as the forget gate f_t^i is close to 1, the information survives to the next iteration. A definition of the variables can be found in the equations 2.1 and the subsequent text.

$$\begin{bmatrix} i_t^l \\ f_t^l \\ o_t^l \\ g_t^l \end{bmatrix} := \begin{pmatrix} sigm \\ sigm \\ sigm \\ tanh \end{pmatrix} T_{4n,2n} \begin{bmatrix} h_t^{l-1} \\ h_{t-1}^l \end{bmatrix} \tag{2.1}$$
$$c_t^l := f_t^l \odot c_{t-1}^l + i_t^l \odot g_t^l$$
$$h_t^l := o_t^l \odot tanh(c_t^l)$$

A visual representation of the above equations can be found in Figure 2.10 with the time steps (t) in the horizontal direction and the layers (l) on the vertical.

The central variables are the memory cell c_t^l and the output h_t^l that stand for the information the cell retains through time and what the rest of the model gets to see from the memory cell respectively. The visible output of a cell, h_t^l, is calculated based on the memory cell as well as a multiplicative output gate o_t^l which is based in turn on the input information h_t^{l-1}. This allows the cell to keep information in the memory cell that will be used for prediction only at a later point in time without this having an impact on the immediate prediction. The memory cell is calculated using previous hidden states: the one of the same layer one time step before and the one of the layer below at the same time step.

Finally, i, f and g stand for the input gate, the forget gate, and the input modulation gate respectively. They control the information flow from the previous state of the memory cell and the previous outputs to the current state of the memory cell.

The functions $sigm$, $tanh$ and $\odot$ are the sigmoid, hyperbolic tangent and product functions respectively and are all applied element-wise.

In the lowest layer (or the only layer in the one-layer case) there is no h_t^{l-1}. Instead the input – this is a horizontal slice of the past data matrix – is fed to the cell.

The prediction is based on the hidden state h_t^l of the topmost layer. As the number of the hidden states does not match the desired output dimension, which is N_S, a trainable matrix that is used to overcome this obstacle.

The network only ever directly predicts one time step at a time. For predictions further into the future, predictions from the preceding steps are used as input values. The prediction matrix is thus built row by row.

Uses of LSTM

Those architectures have been tested and have been found to be superior to traditional methods and feedforward neural networks. However, they are either univariate (Fu, Zhang and Li 2016; Ma et al. 2015; Shao and Soong 2016; Tian and Pan 2015), use small nets (Ma et al. 2015; Shao and Soong 2016; Tian and Pan 2015), only predict in the very short-term or have separate nets for different forecast windows (Fu, Zhang and Li 2016; Shao and Soong 2016; Tian and Pan 2015; Zhao et al. 2017), or use very little training data (Fu, Zhang and Li 2016; Ma et al. 2015; Shao and Soong 2016).

Limited data and weight sharing

Neural networks have been found to be very "data hungry" (Vlahogianni, Golias and Karlaftis 2004), as the expressiveness of neural nets increases with the number of layers and neurons (Goodfellow, Bengio and Courville 2016). One of the challenges, especially for larger scale multivariate predictions, is to find good schemes for parameter reuse. In computer vision, convolutions are a good way of achieving exactly that. Having connections from the entire image to a (large) hidden layer would be expensive. Instead, only small connections between small patches of the underlying data to some hidden layer are learned. The entire image is then "scanned" by those patches, but always reusing the same parameters for every patch, leading to a significant decrease in the total number of parameters that have to be learned. This still requires a large number of calculations, but the number of parameters is typically of greater concern than the number of calculation steps.

In images, the neighbourhood in the data is meaningful: A rectangular subset of pixels showing a stark contrast form a *boundary* of sorts no matter where in the image they are. This border can then be used in higher order semantic reasoning. In traffic on the other hand, geographic neighbourhood is not necessarily that expressive, as the traffic network can be viewed as topological. In certain cases, for example where all stations are on a straight line with the same semantics (no on-ramps mixed with highway sensors), methods that use convolutions can be beneficial (Wu and Tan 2016). Other approaches literally transform the traffic conditions into an image, thus losing the topological information of the street network (Yu et al. 2017). A promising approach merging the convolutional idea with the topological nature of the traffic network was developed by Shahsavari and Abbeel (2015) and used by Yu, Yin and Zhu (2017). The aspects that are promising there revolve around the fact that upstream stations are good (short-term) predictors of traffic flows, no matter where in the network they appear. This relationship does not have to be learned separately for every position in the network. On the other hand, as the street network is not purely topological, it is important to carefully provide information about the idiosyncrasies of the particular situation, such as the distance(s) to the upstream station(s). In addition, relying on convolutions on the graph does not provide global information that could be useful in predictions.

For the most part, deep learning based traffic predictions have been implementations of straightforward end-to-end learning without covariates. Neural networks have been known to be very data intensive (Vlahogianni, Golias and Karlaftis 2004), a requirement that is not easily fulfilled in traffic predictions. The number of days which can be used before a change can alter (parts of) the system is limited to a couple of thousand at best, leading to at most a couple of hundred thousand (highly correlated) training time steps. Finer resolutions of time will increase this number but not the underlying traffic situations that could be observed.

In this situation, it may be beneficial to not only adapt new trends from the machine learning community but to try to incorporate domain knowledge into existing architectures to improve the predictions by making the problem easier for the deep networks to learn.

2.4 Research gaps

Based on this overview over the important concepts in the field and some relevant research results, this section will flesh out the research aims given in Section 1.3 in view of the research that has already been conducted.

The identified gaps form an arc over the analysis of human movement from passively tracked data with large sample sizes. The first is concerned solely with the geometry of the movement. This is followed by a discussion of ways to extract semantics from the geometry and, finally, an application of knowing the semantics is investigated.

2.4.1 Reconstructing geometry from CDR

Section 2.1.3 has argued that passive tracking is a source that will see increasing exploration in the future and Section 2.1.2 has argued that many semantic abstractions of movement rely rather heavily on domain knowledge to obtain their semantics, restricting the possibilities for findings to those that comply with these assumptions.

It is therefore important to explore the extent to which the semantic assumptions can be dropped when working with CDR data. The main challenge with this type of data lies in their temporal sparsity, as mentioned in Section 2.1.3. This sparsity makes it necessary to determine what happened between the (few) points in time where information about the location of a person is available. This leads to the overall aim of this first case study:

> RESEARCH QUESTION 1:
> How accurately can the movement geometry be extracted from call detail records using as few semantic assumptions as possible?

This goal is to be achieved without relying too heavily on semantic assumptions. Those assumptions would limit the applicability to cases where they are met. However, the advantage of very broad coverage would be that most people could be recorded, even those with unusual working hours, multiple homes, or non-fixed working places, where the assumptions may not be fulfilled. Therefore relying heavily on assumptions and rule-based inference could limit the advantage that broad tracking allows, which would not be desirable.

To achieve this, the cell sizes that can vary by orders of magnitude must be considered carefully. Although ping-pong patterns and load balancing can occur both in urban and rural areas, the effect this has on the surmised position of the individual is vastly different depending on the density of the masts. A prior analysis of how this affects the patterns observed is clearly needed.

2.4.2 Mode detection from CSD like data

As can be seen in Chapter 4, there are limits to CDR based localisation that are unavoidable given the size of the cells. Consequently it can not be assumed that the methods mentioned in Section 2.2 could be successfully applied on ID-based data.

However, given that the precision of passive tracking is not limited to cell ID's but can be increased substantially by means described in Section 2.1.3, the question remains open as to whether position estimates that can be obtained now by mobile phone companies are sufficient to extract detailed semantic information on movement based on geometries from passive tracking. The overall question to be answered is thus

> **RESEARCH QUESTION 2:**
> How much worse than GNSS data can passively tracked data be in terms of spatial accuracy and temporal granularity while maintaining the distinguishability of transportation modes?

Because the literature is inconclusive regarding the best methods to be used, even for GNSS based studies, a whole range of different methods will need to be applied to get an impression of a reasonable range in which the expectations could lie.

Because a comparison to GNSS is the aim of this case study, beginning with GNSS data and distorting it seemed like the natural choice, especially as since the (few) datasets with CSD do not usually contain either labels or ground truth positions.

2.4.3 Predicting single mode traffic flows

As shown in Chapter 4, in some cases mode detection on passively tracked CSD can already be an option. It can be assumed that the quality of the mode detection will improve, as technology already developed but not yet applied on a large scale (cf. Section 2.1.3) becomes commonplace.

This allows a more wide spread usage of methods that currently rely on fixed infrastructure such as networks of loop detectors. In particular the geographic scope could be extended, as there is no extra cost for the maintenance of a sensor network that is operated by mobile phone companies anyway. Furthermore, also the semantic scope could also be expanded, as multimodal tracking and changes between modes can also be monitored if the collected tracks are Langrangian rather than Eulerian.

As discussed in Section 2.3.4, the best predictors of traffic flows in the short run are currently based on deep learning. However, as explained previously, the typical paradigm of larger models and more data cannot easily be adapted to the situation of human mobility. More data in this case would mean longer training periods in which the temporal relationships between flows at different locations could be learned. However, cities evolve, new infrastructure gets built, certain areas can go out of fashion, or their use could change, all of which reduces the usability of data from the distant past. Therefore there is a limit to how much history can sensibly be used for learning about mobility in cities.

Additionally, working with limited training data has the advantage of being able to adapt relatively quickly to changed environments if, for example, new lines of public transportation become available.

One goal of deep learning based prediction schemes therefore has to be to simplify the problem as much as possible and to incorporate domain knowledge most effectively, both of which have received only little attention in the past by the traffic literature. This is of particular urgency as the impending use of passive tracking will begin with no history and can be applied in situations that change faster than the road infrastructure, reducing the available history. The research gap to be addressed is therefore:

> **Research question 3:**
> How and by how much can the error in deep learning based traffic flow prediction be reduced by reframing the prediction problem and reducing its complexity?

This situation is different from that of the first case study: In the first case study a priori domain knowledge was not particularly useful in getting close to the ground truth positions, as the focus was on the individual trajectory. However, traffic flow prediction is only interested in sums of flows and not the individual. Therefore including knowledge about how people can generally be expected to behave can turn out to be beneficial.

Chapter 3

Reconstructing geometry from CDR

The end was contained in the beginning.
— Winston in 1984

This chapter, is based on a research article that was published by the *Journal of Location Based Services* and is available at https://doi.org/10.1080/17489725.2017.1333638
The scientific contributions of other researchers to this chapter were the following: Robert Weibel provided supervision during development and helped prepare the manuscript of the research article. Rein Ahas provided the data he had collected for his own research and proofread the manuscript of the research article. Erki Saluveer proofread the manuscript for the research article.

3.1 Study setup

If the extraction of detailed semantically meaningful information based on passive tracking is to have any hope of succeeding, the geometries it provides must be accurate enough to allow this. This first case study therefore is not concerned about inferring semantics and focusses purely on geometry.

The data that is most widely available, as mentioned in Section 2.1.3, are call detail records, or CDR. Their wide availability makes them a natural starting point to investigate the potential for large scale passive tracking exercises.

Why GPS data?

CDR's however rarely come with labels and so the precise analysis of where the quality of an analysis is satisfactory and where it is not is often difficult. One way of handling this would be to compare on an aggregated scale on e.g. statistics published by the authorities (Calabrese et al. 2013; Janzen et al. 2016) or to make more qualitative statements by comparing different regions (Kung et al. 2014). In this case study however, the option to collect ground truth was chosen, as for a downstream semantic analysis, the individual results can be important and not just the aggregated ones. In addition to using individualised data for this study two issues were important to avoid:

Parsimonious assumptions

First , in the inference on CDR like data, people often make semantic assumptions, such as people working during daytime, which may not be justified for the entire large scope of people on which CDR data are available. Many other studies have in contrast readily made use of domain knowledge to extract semantic information directly from the data, without reconstructing the geometry of movement (Ahas et al. 2010; Becker et al. 2013; Isaacman et al. 2011). In this case study on the other hand as few a priori assumptions as possible are made.

Second, the scope should be preserved wherever possible. Many studies, such as Widhalm et al. (2015) have shown impressive results using temporally dense data. However, many people do not generate this abundance of CDR points, as Tanahashi et al. (2012) showed. The straightforward way of dealing with this problem is to not consider those people who do not generate high enough numbers of CDR's (Ahas et al. 2010; Becker et al. 2011; Zhao et al. 2016) or with data that inherently looks more similar to GPS data (Calabrese et al. 2013; Doyle et al. 2014; Schulz, Bothe and Korner 2012). As there may be bias when analysing only the people calling and texting a lot, it makes sense to attempt to incorporate as many people as possible in the analysis, thus the methods applied should be capable of handling most cases, which in CDR means also mobile phone users with only moderate amounts of generated CDR's.

Parts of the case study

To see how well with those two aims in mind the geometry of the movement can be reconstructed, the following steps are performed:

- A representation of the individual space usage that is amenable to identifying repetitive patterns is developed along with a method to infer the missing parts of a day.
- An alternative method for the inference that is based on the literature is also applied to the data for comparison.
- The experiment is repeated in a simulation study, with close control over all parameters.

3.2 Data and preprocessing

Both a real life dataset that shows the behaviour of the methods on messy human data as well as a simulated dataset that allows for more control over certain parameters was used. Section 3.2.1 will describe the real world dataset where as the simulation is described in Section 3.2.3.

3.2.1 Real world dataset

The real life data was used to test the methods comes from two sources: A cell phone app and volunteered geographic information.

YouSense data

The first data set comprises information that was gathered from about 140 Estonian participants during 2015 using the YouSense[1] application first presented in Linnap and Rice (2014) that was since developed further. The data comprises information on 22,943 days of the users, an average of well above half a year per user. The participants were recruited in the environment of the university of Tartu and incentivised by being given a new mobile phone that they could keep after completion of the study. As such, the participants are not representative of any population and generalisations of their movement patterns to the population at large are hardly possible. The collected information includes GPS positions, timestamps of sent and received text messages and calls as well as the connectivity status of the phone (i.e. what mast the phone was connected to at any given point in time).

There are three layers of information contained in the YouSense data:

- The GPS information comes at a sampling rate of mostly one minute, which is adequate as GPS information is only used for evaluation purposes in this study. Three reasons can lead to diversions from the usual sampling rate: The users were allowed to pause GPS recording temporarily, bad reception can prevent a clear GPS signal, and the app can pause recording if the phone does not move.

- The application stores information about the status of the connection whenever that status changes (connected, flight mode, emergency calls only, no connection) along with the ID of the cell (if connected). This information will be referred to as handover data even if strictly speaking it is a bit richer due to the information beyond the simple cell ID.

- The CDR are annotated with a time stamp and and the nature of the record (e.g. incoming call, outgoing text message).

[1] http://positioner.ut.ee/dashboard/info/

OpenCellID

In order to bring together the GPS coordinates with the information on the cell towers the information from a second source was used, namely OpenCellID[2], which constitute volunteered geographic information (VGI) (Goodchild 2007). The information contained there is recorded by volunteers that run an app that records both GPS positions and the cell ID's. This information is then uploaded and aggregated.

The geometric midpoint of all the GPS positions is taken as the one point that the cell is represented as. This has the advantage that the position of antenna is irrelevant, as only the actual positions of people connected to the cell are relevant. The midpoint of all positions of people connected to the cell is a reasonable cell ID based proxy for positioning.

Taking the geometric midpoint makes this estimate sensitive to outliers and duplicate cell ID's. In particular since some cells, particularly in rural areas are not recorded too often, a single wrong location can shift the midpoint considerably. Duplicate ID's on the other hand can lead to two locations on opposite sides of the country can be merged into a position that makes no sense for either of the locations.

A summary of the different kinds of data that was used can be found in Table 3.1.

Content	Source	Use	Description
GPS	YouSense	Ground Truth	Sampled at most once per minute.
Handover data	YouSense	"Ground Truth" at cell granularity	Connection to all the masts (even if no CDR is produced). Indication if no connection was possible
CDR	YouSense	Input for the extraction of typical days	Time and type of all CDR activities.
Cell locations	OpenCellID	Connect IDs and locations	VGI

Table 3.1: Summary of the data used in this case study and its use.

Preprocessing

The data was slightly pre-processed the GPS to allow for easier and more reasonable comparisons: If the distance between two GPS recordings was large in terms of space or time (500m and 5min respectively) or if the temporal distance was very large (greater than two days) the time between them was disregarded. Next the trajectories were smoothed where they showed indoor behaviour and flagged as stop every fix which has neighbouring fixes in a contiguous time interval of at least five minutes in which there is no GPS signal outside a 100 m radius around the measurement. The next step flags hitherto unflagged points if the containing sequence of unflagged points is "short" (thresholds based on total distance travelled, total time, circle radius and number of points). The segments were then sequences of points with the same flag status.

[2] www.opencellid.org

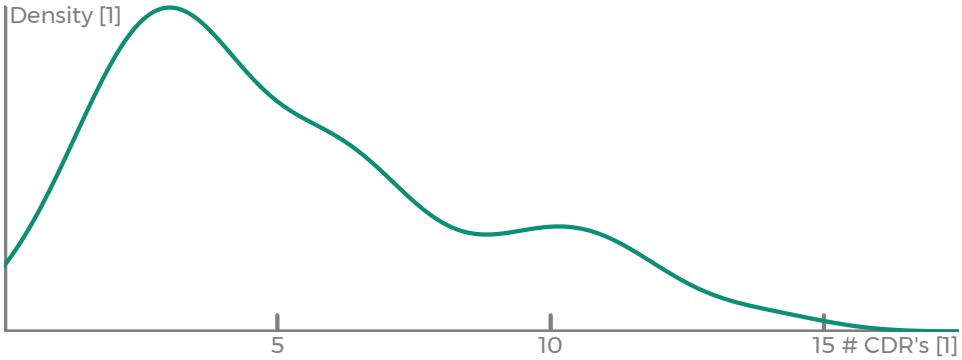


Figure 3.1:
Probability density function of the average number of daily CDR's per user.

A first informal look at the data revealed that the number of people without a clear work location and/or irregular movement behaviour was larger than expected. In the context of the current work, this is not detrimental to the case study, as the methods are expected to work well even for users with unconventional yet regular behaviours.

3.2.2 Exploratory data analysis

In this section different aspects of the data are visualised in order to provide the reader with an intuition about them. This should facilitate the understanding of why some of the choices were made the way they were.

Summary statistics on the data

Daily CDR counts and intervals

A very important question is the number of CDR that can be observed for a typical day of a user. Even if the number is high, this does not mean that there is good coverage of an entire day, as the CDR may be temporally intensely clustered. If there are just few CDR's however, there is certainty that temporal coverage is lacking.

While point measures are inadequate to describe the whole behaviour, averages by user shown in Figure 3.1 can give a first idea of the order of magnitude. As can be seen, most users have something between three and four CDR's per day on average, with some users being significantly more active at daily CDR counts of above ten.

To investigate the distribution on a high level, it can be instructive to look at the time that passes between CDR's. If the distribution of those time differences is concentrated then the CDR are spread widely throughout the day and a broad view on the movement of a person may be possible even if there are not that many of them.

In the YouSense data however, as the peak at very low values in Figure 3.2 shows, CDR's tend to happen in bursts, reaffirming findings from the literature on CDR's (Jiang et al. 2013b) and human behaviour in general (Barabasi 2005). For a fixed total number of CDR's, the concentration of information on location on short temporal intervals makes localisation more difficult the rest of the time.

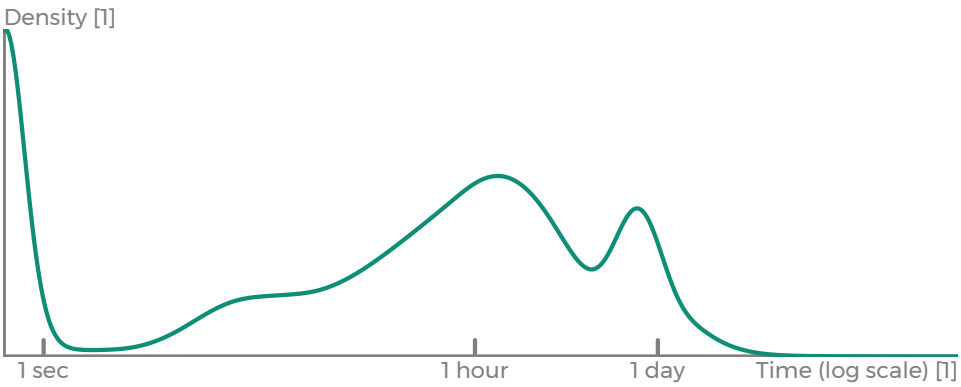


Figure 3.2:
Probability density function of the time difference between consecutive CDR's using a logarithmic time scale. While the abscissa is without unit, the labels for the ticks have been translated back to be better understandable.

Table 3.2:
Radii of gyration (ROG) and distance travelled in km on the three available levels of the data: GPS, Handover (GSM) and CDR. Superscripts explained in the text.

Statistic	GPS	GSM	GPS^S	$GPS^{W,S}$	CDR	CDR^W
ROG	10.2	10.4	10.9	9.9	6.2	6.8
Dist. travelled	52.6	221.3	35.6	-	23.1	-

Given those low numbers for CDR's per day, CDR's should not be looked at as sparsely sampled trajectories. To present the effect that such a view would have, some statistics, namely the radius of gyration and an estimation of the distance travelled that are also used in the literature (Ranjan et al. 2012; Schulz, Bothe and Korner 2012) will be shown next.

Distance travelled and radius of gyration

For this the definitions the *distance travelled* and both the *weighted* and *unweighted radius of gyration* (ROG) are needed. Both will be calculated for every user day and then averaged to produce a single number.

The distance travelled is defined as if the movement between two fixes were performed in a straight line. This is asymptotically true if the sampling rate is high enough but cannot be expected to hold in the case of the YouSense data.

$$\text{RoG} := \sqrt{\frac{\sum_{i=1}^{N} w_i \cdot (p_i - \bar{p})^2}{\sum_{i=1}^{N} w_i}} \qquad \text{with} \qquad \bar{p} := \frac{\sum_{i_1}^{N} w_i \cdot p_i}{\sum_{i=1}^{N} w_i}$$

with N the number of points and w_i the time spent in p_i for the weighted version and $w_i \equiv 1$ for the unweighted version. As the time spent in p_i the time to the next fix is assumed; an approximation that decreases in quality with sparser and sparser sampling rates.

The results for different layers of information in the YouSense data can be found in Table 3.2.

The following comparisons are made: GPS trajectories (GPS), the handover data seen as trajectories (GSM) and the CDR data (CDR) are taken as trajectory the way the were recorded without modifications. In addition, variations thereof focusing on the movement between stops (S) and time weighted versions of the summary statistics (W) are also calculated. The reasoning behind the different numbers is the following: GPS^S attempts to bridge the gap between GPS and GSM in that the stops of the trajectory are extracted (using spatial and temporal thresholds) and the movement is assumed to be none during the stops and instantaneous between them. This can be assumed to be closer to GSM assuming most calls happen outside movement segments. $\text{GPS}^{W,S}$ then uses the time spent on the stationary parts of the trajectory as weights for the ROG calculation. CDR^W essentially treats every CDR point as a stay in the sense of GPS^S that lasted from the temproal midpoint to the previous to the temporal midpoint to the subsequent observed point. As both time-weighted schemes only affect ROG, their distance travelled does not change from the unweighted version and therefore is not reported in the table.

Due to load balancing and ping-pong effects the active cell can change even for the large parts of the day when a typical person does not move. Therefore it is not surprising that GSM clearly overestimates distance travelled. However, it its estimation of the ROG is not too bad, indicating that the essential aspects of the movement as measured by the radius of gyration could have been captured.

Due to the typically small amount of time spent on move segments, the ROG of GPS^S are close to the GPS radii. On the other hand the distance travelled as measured by GPS^S is below GPS, as the assumption of movement in a straight line between stops is clearly a simplification.

CDR and CDR^W both cannot capture movement that happened in between CDR's. Therefore it cannot come as a surprise that the distance travelled is significantly underestimated.

Pairs plot of ROG

Instead of averaging all user days in a single number, producing pairs plots can reveal more of the connection between the different numbers. Figure 3.3 demonstrates the correlation between the different radii of gyration. The correlation between GPS and handover are fairly high with almost 99%. In addition, those correlations do not depend on the amount of information available for those days (as indicated by the colours), which confirms expectations. Note that the GSM values are more dispersed, as the individual positioning of the cell centroids influences the result. The two measures based on CDR are markedly less correlated with GPS measurements 83%. Clearly visible are the days with only a single cell ID where the movement and thus radius of gyration is estimated as zero. The fact that in many of those cases actual movement is significantly above zero means that the CDR's were either just very few in number or temporally highly clustered.

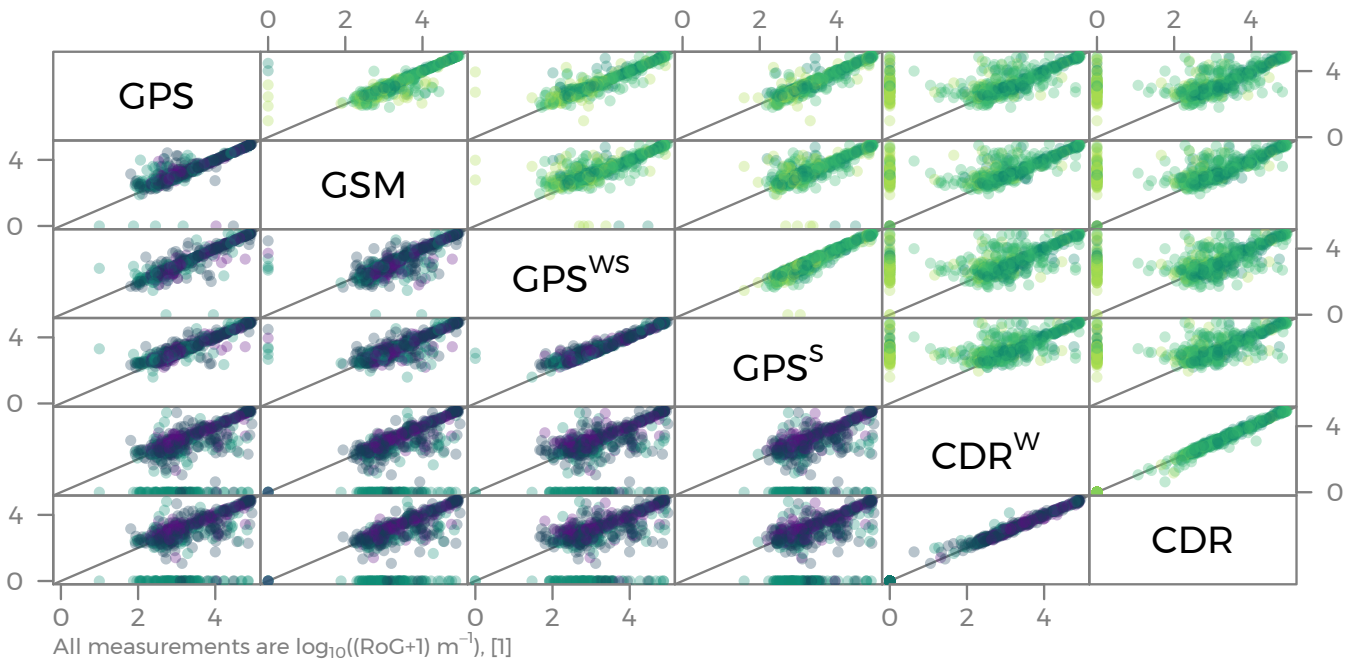


Figure 3.3: Pairs plot of the radii of gyration, measured in different ways explained in the text and indicated on the diagonal. The points represent 600 randomly selected days from the database for which there were at least two stop points. The pairs plot is symmetrical in the positioning of points, yet not in colouring. Bright colours indicates the quartile with the least information, dark colours indicate the respective quartiles with the most information. Below the diagonal the amount of information is measured as the time-weighted average fraction of the day that a CDR is closest; above the diagonal, information is measured in CDR counts for that day. Some jitter was added to the locations to reveal areas of high density.

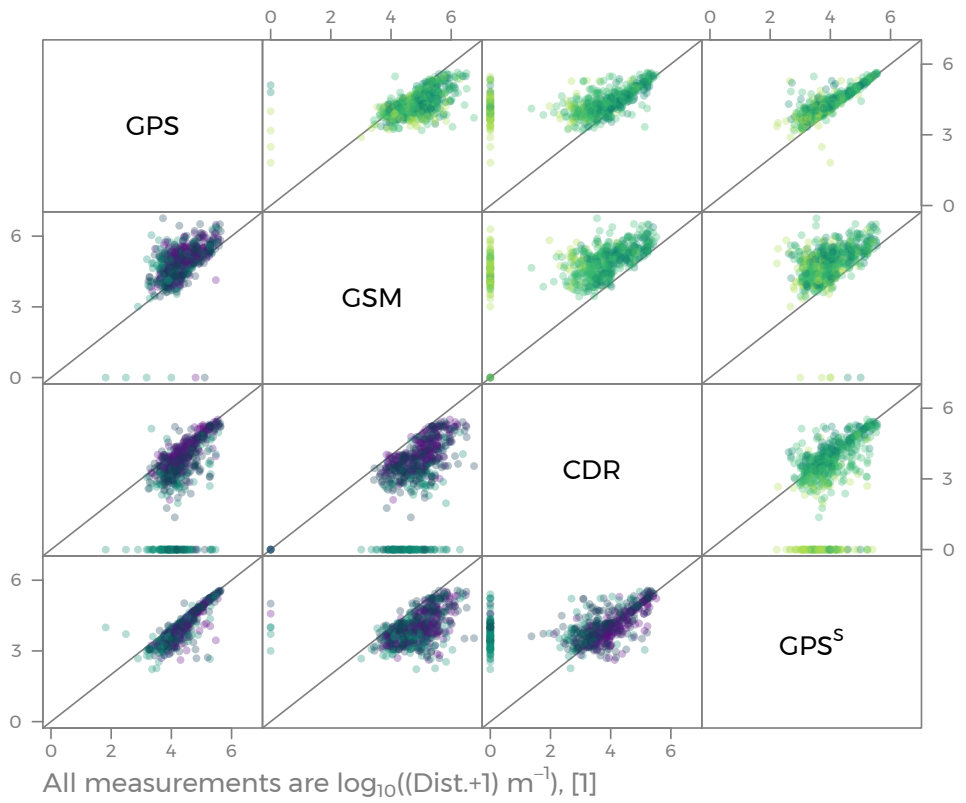


Figure 3.4:
Pairs plot of the daily travelled distances, measured in different ways explained in the text and indicated on the diagonal. The points represent 600 randomly selected days from the database for which there at least two stop points. The pairs plot is symmetrical in the positioning of points, yet not in colouring. Bright colours indicates the quartile with the least information, dark colours indicate the respective quartiles with the most information. Below the diagonal the amount of information is measured as the time-weighted average fraction of the day that a CDR is closest; above the diagonal, information is measured in CDR counts for that day. Some jitter was added to the locations to reveal areas of high density.

Pairs plot of distances

The other statistic that was calculated were the daily distances travelled. The corresponding pairs plot is Figure 3.4. Here it can be seen that the GSM distance is mostly higher than the GPS distance. This is the result of fictional movement induced by changes in active cell without underlying movement of the person. An additional contributor to the overestimation of movement using handover data is the fact that all movement is assumed from cell mid-point to cell mid-point. This can result in a much more tortuous movement implied by the handover data than took place in actuality. There is an antagonising effect that movement while connected to the same cell is underreported, but judging from the figure, this effect is of much lower importance than the aforementioned effects. Thus the correlation is markedly lower than for the radii of gyration and is only 47%.

The over-reporting by handover data is also observable – and even stronger – if the comparison is with the stops only GPS trajectories. This is because the antagonising effect is smaller due to omitting movement during a stop, where the GPS signal typically implies movement even if actual movement did not take place. The correlation there drops slightly below 46%.

In the same vein the CDR distances are naturally lower than those based on GSM, as the fixes used in the former data source are a subset of the fixes of the latter. Interestingly the distances between GPS stops are relatively close to the distances from CDR, at least for the days with more CDR information, but naturally still very big on days with few CDR.

There seems to be a tendency for the CDR based measures to get closer to the GSM based numbers as the information increases, which is also what one would expect: In the extreme case of CDR's every second, the two measures should coincide.

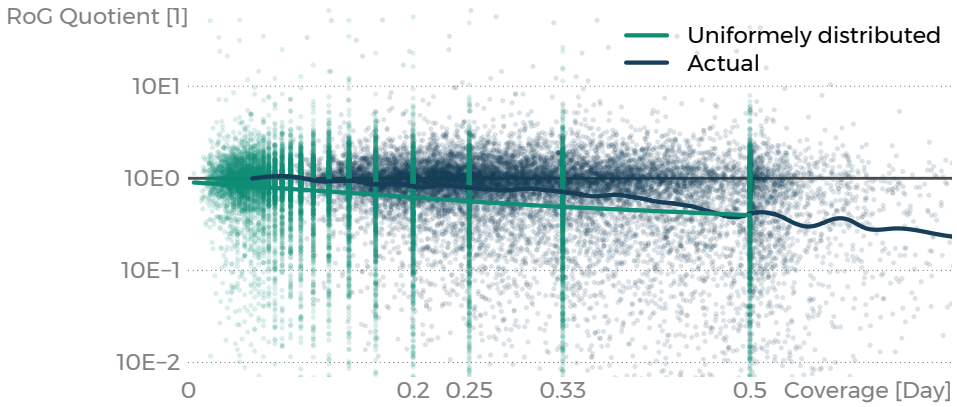


Figure 3.5:
Quotient of RoG calculated from the CDR-"trajectory" and the RoGs based on GSM. In blue with average observed fractions of the day covered by a CDR of that day, in turquoise with one over the number of CDR's that day as the average fraction. Smoothing splines for both solutions (on the log scale) in darker colors.

Quotients of RoG and distance

This tendency can be confirmed by looking at the RoG based on CDR and based on handover data. In order to visualise this, the quotient between the two RoG have been plotted against the average fraction of the day that a specific CDR determined the position. This fraction can either be calculated as one divided by the number of CDR's on that day or by taking a weighted average, weighting by the duration for which a CDR determined the position. The shorter that duration (i.e. the more CDR there were), the closer to 1 the quotient gets meaning that the two RoG approach one another. The visualisation is drawn as Figure 3.5.

While the RoG based on CDR can be close to the one based on GSM data for high enough numbers of CDR, the same does not hold true of the distances. Every missed cell reduces the distance, and as the number of connected cells during a day is typically large, there is simply no hope of getting even close to the GSM distances with the numbers of available CDR's. Interestingly however the quotient of the distance measures does not deteriorate with lower values of CDR if the distances based on GPS are in the denominator, as illustrated in Figure 3.6.

The figures also show the differences of typical time per CDR based on the two ways of calculating it. The fact that the points in turquoise are to the left of the points in blue is reminiscent of the bursty nature of CDR (Barabasi 2005; Gonzalez, Hidalgo and Barabási 2008; Song et al. 2010). The vertical lines of turquoise points are at the values of $1/N$, whereas the blue points are at the time weighted actual average fractions where a CDR point was determining position. The first way of calculation underestimates the time for which a CDR is the closest, as seen from the Figures 3.6 and 3.5. A natural lower bound of actual (weighted) average fractions of a day covered by a CDR is the square of the fraction of the day spent sleeping (corresponding to zero time between CDR's while awake). For a sleep duration of eight hours this corresponds to 0.11, which is about the minimum of what can be observed on the actual values in blue. While passive (incoming) CDR's can be received during sleep, the bound is nonetheless an interesting value to have as a comparison.

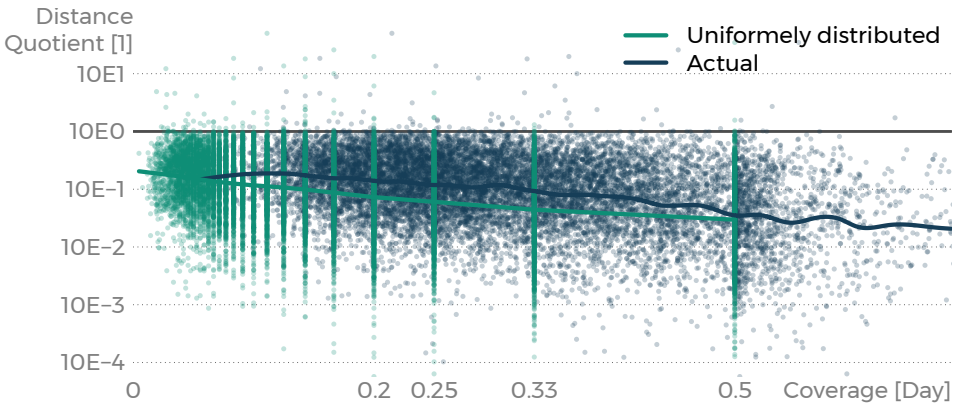


Figure 3.6:
Quotient of daily distances calculated from the CDR-"trajectory" and the distances based on GSM. In blue with average observed fractions of the day covered by a CDR of that day, in turquoise with one over the number of CDR's that day as the average fraction. Smoothing splines for both solutions (on the log scale) in darker colors.

Every step from GPS to handover data to CDR data marks a strong deterioration in locational accuracy.

From CDR to handover the disadvantages of taking information at the cell level are revealed: The movement is overestimated, as technical procedures such as load balance imply non-existing movement. However, certain summary statistics of the movement such as the radius of gyration can be captured with handover data.

From handover data to CDR, the temporal sparsity becomes evident. While the overestimation of distance travelled decreases sharply (sometimes to zero on days with few CDR's), the radius of gyration still is highly correlated to the one based on GPS for days that have CDR's in more than one location.

User days visualisation

Now that an intuition about the frequency and distribution of the data on the different is established, a look at the movement itself can be insightful. The visualisation in Figure 3.7 is an overlay of the CDR's over the stops identified through GPS. For this the stop segments from the segmentation were clustered (DBSCAN with $\epsilon = 30$m and a minimal number of 4 points) and then classified according the time when those stop locations were visited (which admittedly is a very crude proxy for a semantic layer, but it helps in differentiating visually the different stops). Places of overnight stays are shown in hues of blue, places visited for more than three hours a day on average are shown in reds and other frequent places are shown in green tones. For each of the color scales, the most visited n ($n = 2$ for blue and red locations and $n = 6$ for green locations) places have their own colour, such that e.g. all dark red rectangles correspond to visits to the same location. Further locations of the same type share the same colour, so that light blue locations need not be the same. Infrequently visited locations are shown in yellow and movement segments are shown in gray. The CDR are grouped into active (outgoing calls and sms, represented as red triangles) and passive CDR (incoming calls and sms, represented as orange crosses). The light gray bars in the background represent the weekends.

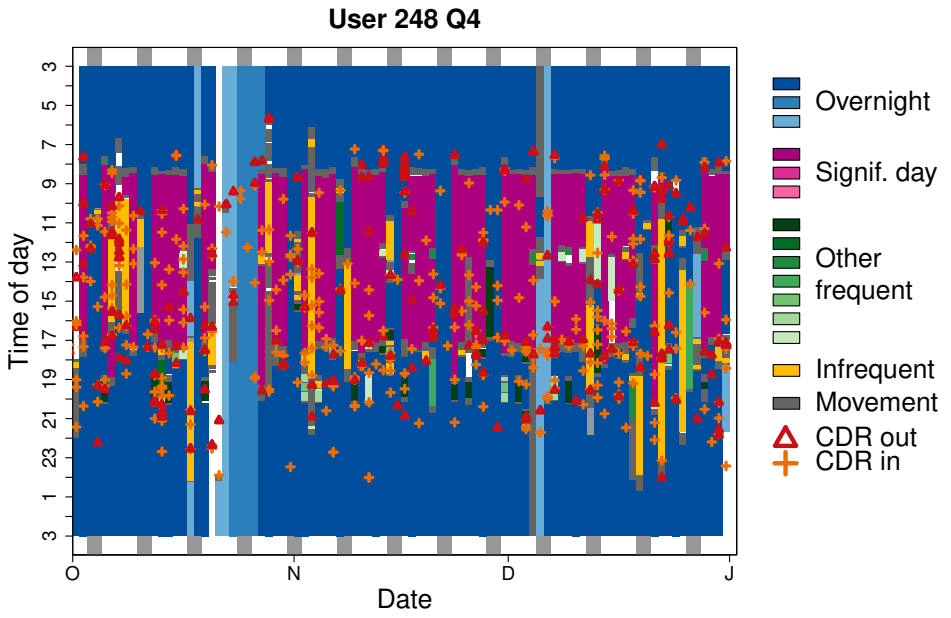


Figure 3.7:
Visualisation of CDR data (foreground) vis-à-vis the (interpreted) GPS data in the background. For readability's sake the plots are drawn for every user and every quarter of the year separately. This example is the final quarter for user 248.

The user in this example has a relatively stable workday-weekend structure and CDR's that are spread evenly over the usual waking hours one expects, with a slight over-representation of the times from 5 pm to 9 pm. In addition there seems to be a week in the end of October and mid December where the usual pattern is interrupted by irregular behaviour. The white day on the third Tuesday of October is a day where the user paused the recording of GPS data (though not CDR data) and therefore for that day no information is available.

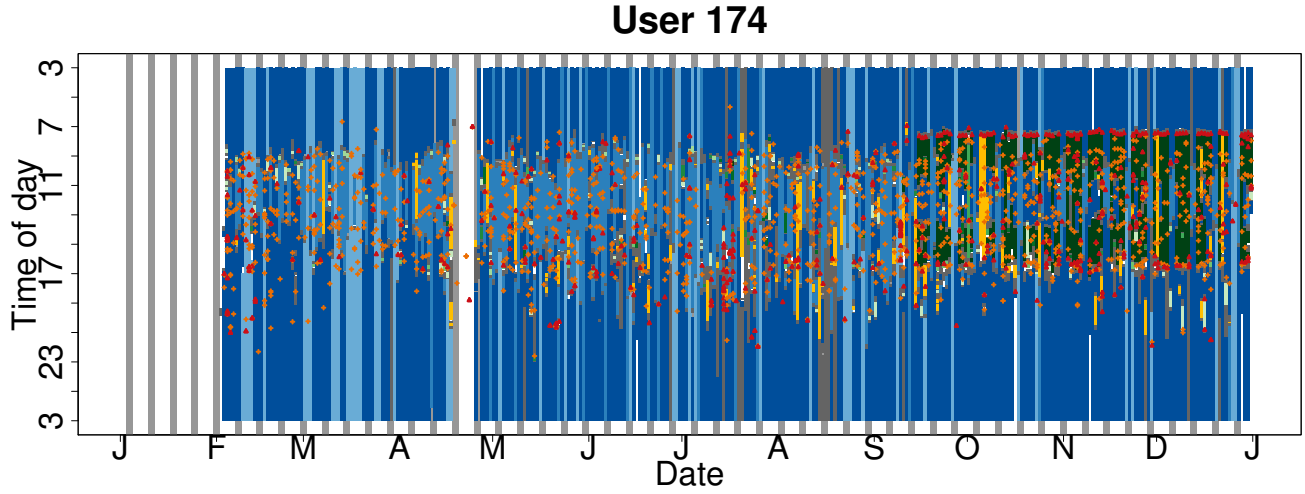
Visualisations of different users

In order to demonstrate the diversity of behaviour that could be observed in the data, similar visualisations for three different users are available as Figure 3.8. The entire period for which data was available is shown for every user. This has the disadvantage that the temporal scales of the subfigures are not aligned, but what is gained is some breadth of the daily bars, making the plot more readable.

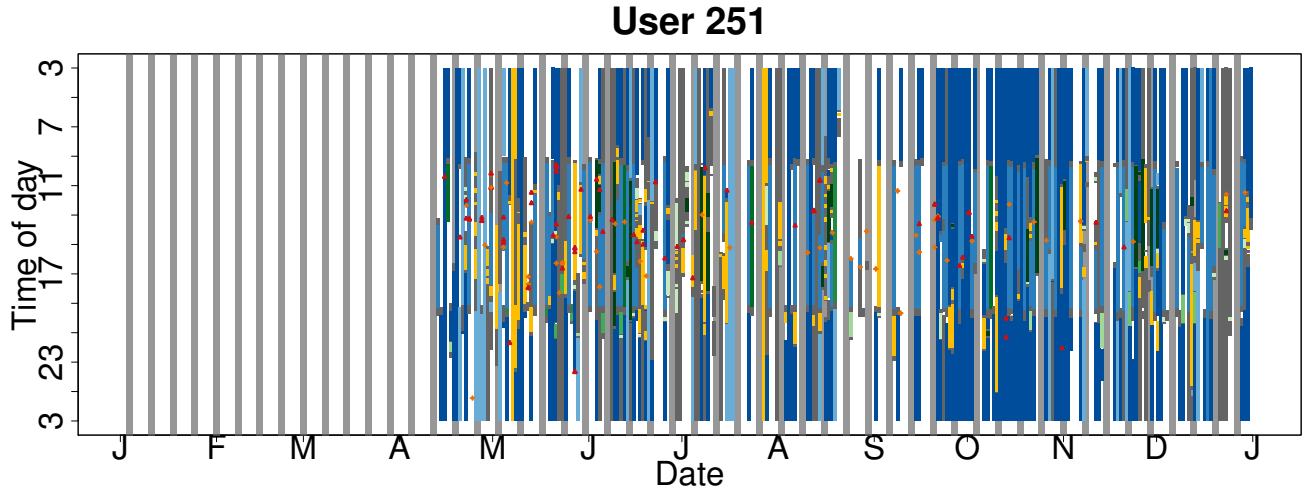
Clealy the regime change of User 174 in September is visible where the main daytime location as well as the duration with which the user stays at that location changes. The regularity with which there is CDR activity at nearly the same time every day in the morning is unique to the dataset. It is a good example of CDR's not being able to capture all movement: Note that on no day there is a CDR before the green activity, which makes this location invisible to all analyses that rely solely on CDR's.

User 251 produced unusually few CDR. Later on, when the user days based on CDR's only are visualised, this will become blatantly obvious. Also note that there are many days on which there is no GPS reception, indicating either no reception or a pause called for by the user.

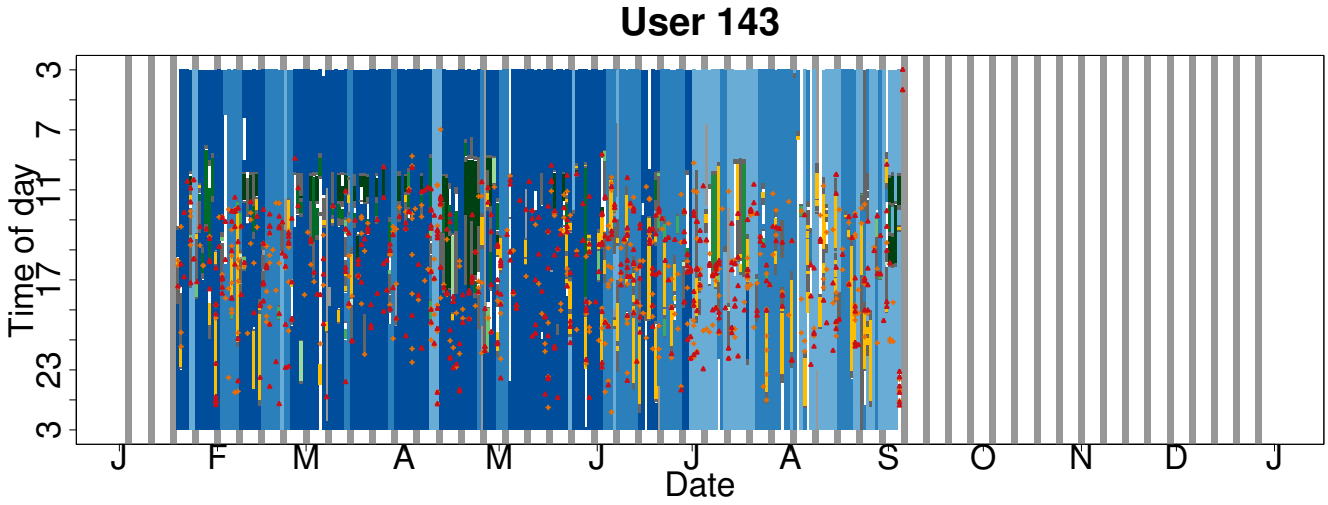
User 143 does not conform to usual expectations at all. First there is a wide variety of locations at which the person spends their nights, which would cause problem if at some point one were to make the assumption of the existence of *a home location*. In addition there is no clear mein daytime location that would be visible and the visited locations are very diverse, as indicated by the many locations that are visited just rarely.



(a) User 174



(b) User 251



(c) User 143

Figure 3.8: Visualisations of the GPS signals and the CDR for different users over the whole study period.

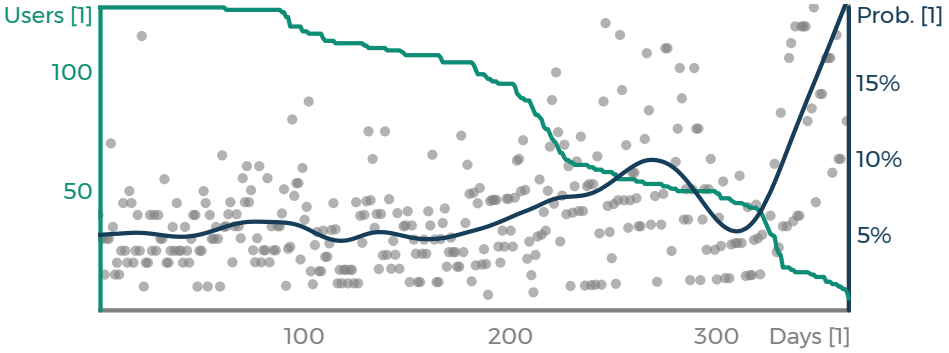


Figure 3.9:
Probabilities to pause GPS recording against the day of observation. In turquoise one can see the number of observed user days, measured from the beginning of the recording of every individual.

Probability of recording pauses

In the visualisations of the interpreted geometries of the GPS movement, some users showed some gaps in the GPS recording. Those could have different reasons, such as a loss of GPS reception, an empty battery of the recording device or the deliberate pausing of GPS data collection that was permitted.

As long as the reasons remained approximately the same for the entire duration of the analysis, all days can be treated the same way. Too see whether this was cause for concern, Figure 3.9 shows the probability for a user to pause the GPS recording deliberately. The time on the abscissa is relative to the starting date of the recording for every user, in order to reveal possible habituation patterns such as a reduced propensity to pause after a couple of weeks into the collection.

Contrary to expectations, the probability of a user asking for a pause in GPS recording did not decrease in time after the first day of observation. Only at the very end of the observation period, where the number of observed users is low and thus the variance is inherently higher a change in the probability is discernible.

Cell density

In addition to the look into the YouSense data, some information about the OpenCellID dataset may also be useful to visualise. As has been stated many times (e.g. (Kung et al. 2014; Rinzivillo et al. 2012; Steenbruggen, Tranos and Nijkamp 2015)), the masts are more dense in cities than in rural areas. The OpenCellID data conforms to this expectation, as can be seen in Figure 3.10 where the densities of the average distances to the Voronoi neighbours for every mast are shown. The bimodality of the distribution emerges clearly, reflecting the expected division. The LTE masts are a bit sparser than those of the other two connection types, but this does not change the overall picture by much. Any method that uses proximities between cell locations should therefore be robust against differences in densities of about an order of magnitude.

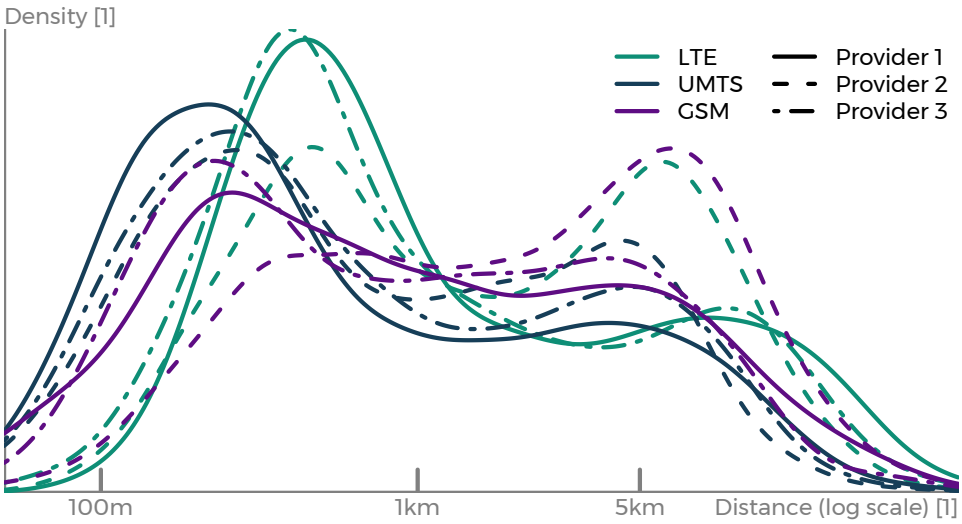


Figure 3.10:
Density of the average distance to Voronoi neighbours by Provider and Connection type. Although the abscissa is without unit, the labels for the ticks have been translated back to be better understandable.

3.2.3 Simulation study

Simulated data can provide a sanity check under messy real world conditions, especially if the sampled population is by no means representative of the population at large. However it can also be helpful to evaluate and compare different methods also in the tightly controlled environment of a simulation to clearly identify circumstances where a method works particularly well or badly. The (regular parts) of the movement of people using the masts from a 10 km radius around Tartu are simulated. To avoid the necessity of first clustering masts, the masts were chosen such that they are at least a certain distance apart from each other (1 km and 500 m respectively, depending on the number of drawn masts). Next, random routines are created for each user, each consisting of a morning, an afternoon and an evening activity with the rest of the time being spent at a randomly drawn home location. Those routines are created directly in vectorised form, such that the upper limit to reach is known to be zero deviation. 200 sample days were created from those routines (the number reflects the real life participants), adding uncertainty about the exact beginning and ending of each of the activities to allow for temporal uncertainty. The underlying routine for every sample day is selected among the available routines for a user with exponentially decaying probabilities in order to reflect the fact that some routines are more frequent than others. Every of those simulated locations is observed with a probability that is proportional to a linear combination of the observed hourly CDR frequencies and a constant probability for every hour.

For every combination of the following parameter choices, 20 users are simulated with 200 days each resulting in 64′000 simulated days:

- Number of locations: Either 5 or 15 masts are used as pool from which to generate routines. 5 is chosen as the upper bound of the very few locations that most users seem to spend most of their time according to (Bayir, Demirbas and Eagle 2010), 15 is a number large enough to allow most routines to happen in (almost) disjunct non-home locations.

- Number of routines: Either 2 or 4 routines are generated. 2 To reflect a Weekday-Weekend dichotomy, and twice as much, to add more complex behaviour.
- Calling probabilities: Either the empirical probabilities for CDR's (EP) or $0.6 \cdot \text{EP} + 0.4 \cdot 1/24$ are taken as base (scaled to sum to 1 over a day). The linear combination was chosen to see whether the first and last locations are fitted better when the CDR's are more dispersed.
- Factors: The base probabilities are then multiplied by 3 or 6, resulting in expected CDR counts of 3 and 6 respectively. The choice for 6 (on average) was taken as half of what (Becker et al. 2013; Isaacman et al. 2011; Pappalardo et al. 2013) used or had, as the focus of this work is on methods that work on moderate counts of daily CDR. That value can then be halved again to see how far down the methods still capture the essence of the movement of a given day.

3.3 Methods

There is a consensus that daily human mobility patterns show a high regularity (Lu et al. 2013; Schneider et al. 2013; Song et al. 2010). A reasonable assumption therefore would seem that this regularity, once learned, should be conducive to the quality of reconstructing the whereabouts of mobile phone users.

This section discusses approaches to use the established repetitiveness of human movement to fill the large gaps between CDR observations. One is based on transactions, without the necessity to consider space, while the other uses space directly.

3.3.1 Representation of the data

As stated previously CDR activity has been found to be "bursty", with a considerable number of CDR's happening in close temporal proximity of others. This can result in an overrepresentation of certain cell-towers in the data.

Day represented as vector

A possible representation to solve this issue can e.g. be found in (Furletti et al. 2012): The day is partitioned into equally sized intervals and the CDR is recorded in the interval in which it happened. As every time slot has the capacity for only one piece of information, only one CDR can be considered; the one closest to the centre of the time slot was chosen. This way, most of the members of the bursts are binned together. Should the burst fall right on the (arbitrary) border of the time slots, both adjacent slots will only contain the same information if there is no other CDR in one of the slots that is closer to the respective midpoint of the slots. The resulting vector is often not complete, as e.g. with two CDR's in a day, at most 2 time slots may be filled.

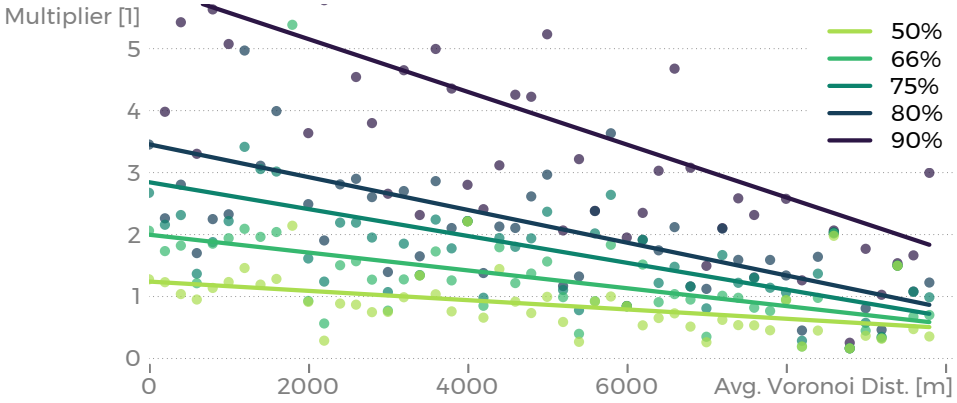


Figure 3.11:
Multiples of the average distance to the Voronoi-neighbours of a cell that describe the necessary radius of a circle around a cell centre required to capture at least a certain quantile of the GPS locations that were recorded while connected to the mast.

The finer the temporal partition, the sparser the resulting data vectors will be populated. There did not seem to be any natural best choice for the temporal granularity so the analysis was performed with partitions of the day into 24, 12 and 6 time slots. Finer partitions seemed to be too granular for the CDR observed counts whereas coarser partitions would correspond to blocks larger than 4 hours which already seem at the border of what seems sensible.

Clustering cell locations

The spatial location is measured on the granularity of the cell regions. Due to the nature of the connections it makes sense to cluster cells that were frequently used and that are close together into a location that is meaningful to the user. The approach chosen by Do and Gatica-Perez (2012) uses equally spaced grids for the analysis and thus does not take into account that the density of the cell towers varies by at least an order of magnitude between urban and rural settings. On the other hand, Csáji et al. (2013) uses twice the maximum distance of a cell to its Voronoi-neighbours to cluster the points.

While there was no information available about how the factor of two was calculated, a look at the data reveals that this factor seems to be dependent on the distances to the Voronoi neighbours, as illustrated in Figure 3.11. The abscissa represents the average distance of one cell location to its Voronoi neighbours. Cells in rural areas that are further apart thus are typically found to the right. The ordinate represents the multiple of the distance to the Voronoi neighbours that is required to capture a given quantile of the GPS points that were recorded while connected to a cell. Values that are high indicate that the GPS locatios of that cell were far away compared to the average distance to the neighbours of a cell.

In addition to performing the analysis with the averages of distances to Voronoi neighbours, it was also performed with the maximum distance to the neighbours instead. Qualitatively the results are the same. As the maximum is more strongly affected by certain anomalies and outliers, the results of the analysis using the averages is shown here.

The multiples required are of course noisy due to the limited number of users, as cells that "see" few users might have their estimation dominated by the distance to the frequent location of a single user. However, there seems to be a clearly discernible linearity in the trend, as the lines drawn are in fact smoothing splines – not regression lines – and could bend if the data suggested non-linearity. The fact that the multiplier should depend on the distance to the Voronoi-neighbour seems natural: While a cell in the inner city may easily serve a phone three cells away due to the high density of the cells, a rural cell with 10 km distance to neighbours may not be able to do so. When measuring distances from cell centres, use scaled versions of those distances were used, i.e. they were divide by the expected radius of the circle containing 75 % of GPS points. For ranges between 50 % and about 80 %, different choices for the threshold scale the adjusted distances approximately linearly, which can be fully compensated by the clustering that will follow, so within this range any value can be chosen, so the value of 75 % was retained.

The rescaled distance matrices between the used cell centres for each user was then calculated and used as input to DBSCAN (Ester et al. 1996) to find clusters. The IDs of cells from those clusters are then changed to the corresponding cluster ID and the location of the cluster are set to the mean of the locations of the contributing cells' centres. Apart from identifying potentially semantically meaningful places of a person, this has the advantage of reducing the number of recorded "cells", facilitating the recognition of patterns.

3.3.2 Reconstructing trajectories

In this section the different methods for reconstructing trajectories from CDR data are presented.

Method 1: Association Mining

The first method is the mining of association rules using the *apriori* algorithm (Hahsler, Grün and Hornik 2005; Ye et al. 2009) with a combination of cell ID and time slot as input. A low support threshold (2 items) was chosen to get broad range of potential rules. Given a sample day with the recorded CDR's in their respective time slots as left hand side the method then looks for the rule with the highest lift for every missing time slot and fills the gaps in this way. Time slots that have no rules given the observations are filled with the closest available information after the rules have been applied. This typically is the case for the very early and very late time slots, that are then simply filled with the first/last predicted location. The advantage of this approach is that it is relatively stable and can deal with different amounts of CDR: the more data it is fed, the more nuanced the rules can become. On the other hand it does not embed any notion of temporal proximity, as the items are just (uninterpreted) labels and the rules it finds are again on a label level and thus the method does not produce representative days by design.

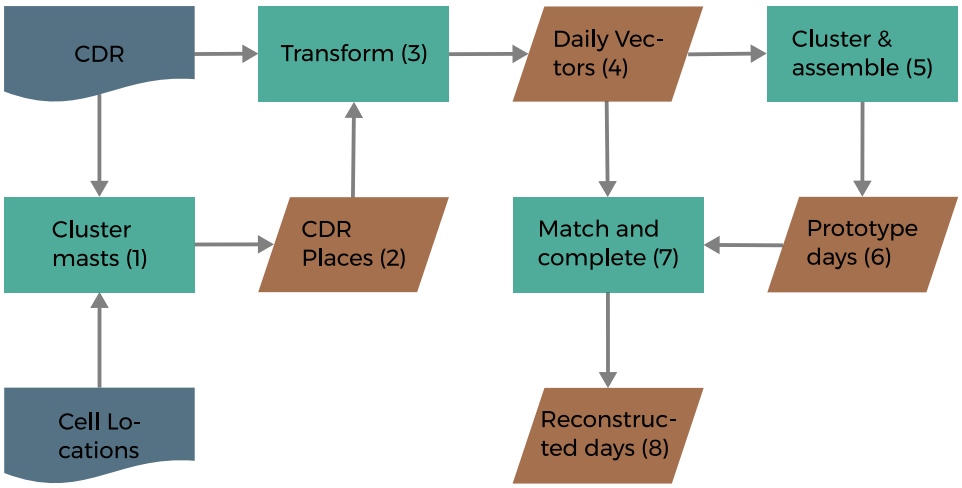


Figure 3.12:
Workflow of the DAMOCLES approach with the enumeration of the steps to be taken.

Method 2: DAMOCLES

The alternative proposed here is DAMOCLES, the *DAily MObility CLustering for Extracting Space usage*. The overall idea is to use the temporal regularity of human movement behaviour to identify typical movement behaviour. In essence this attempts to compensate low CDR counts with aggregation over time. Days that are similar are identified and used to create prototype days, which then can be used to complete sparsely populated daily CDR vectors. As the first method, DAMOCLES is an approach that searches recurring patterns in the existing data. However, in contrast to association mining where all the available items are independent from each other and do not themselves carry spatial or temporal information, it explicitly captures the spatial and temporal structure of the data. There are three main parts to the algorithm, whose schematic representation can be found in Figure 3.12:

1. A dissimilarity measure,
2. a clustering algorithm and
3. a reconstruction based on the identified clusters.

Signatures of the steps in DAMOCLES

More rigorously, and using the step numbering in the graphical representation of the procedure in Figure 3.12, given the set of cells $\mathfrak{C}$ and the number of time slots $n_t \in \mathbb{N}$ that a day is divided into, the extended set of cells, $\mathfrak{C}_e := \mathfrak{C} \cup \{\text{NA}\}$, the set of days $\mathfrak{D} := \mathfrak{C}^{n_t}$ and the set of extended days $\mathfrak{D}_\mathfrak{e} := {\mathfrak{C}_\mathfrak{e}}^{n_t}$ are defined. The extended versions are needed, as the observations may contain missing values. $n_d \in \mathbb{N}$ days $D^o \subset {D_e}^{n_d}$ (Step 4) are then observed. A dissimilarity function $d : D_e \times D_e \to \mathbb{R}_0^+$, a clustering method $c : {D_e}^{n_d} \to \mathbb{N}^{n_d}$ and a cluster assembly method $a : {\mathfrak{D}_\mathfrak{e}}^n \to \mathfrak{D}$ for some $n \in \mathbb{N}$ (Step 5) are then used. Lastly there is a need for a reconstruction $r : {D_e}^{n_d} \times \mathbb{N}^{n_d} \to D^{n_d}$; $D^o \times c(D^o) \mapsto D^r$ where D^r denotes the reconstructed days (Step 6).

Implementation

For the clustering (Step 5), DBSCAN was chosen, as it allows for different numbers of identified clusters per user and can accommodate users with different number of recorded days. Having a method that allows for different numbers of clusters is required as some users might simply have a weekday and a weekend routine whereas others might show more diverse regular days. DBSCAN needs as input a dissimilarity matrix with pairwise dissimilarities between the entities that need to be clustered. This dissimilarity matrix is calculated using d, which needs to fulfil positive semi-definiteness and symmetry in order for DBSCAN to yield sensible results. Note that it does not need to be a metric, as DBSCAN can cope with d fulfilling neither the triangle inequality nor the identity of indiscernibles. Special care should be given to how the dissimilarity treats the missing values: One has to avoid DBSCAN connecting everything through (almost) empty observations. Therefore:

Distance between days

$$
\begin{aligned}
&d(day^{(1)}, day^{(2)}) \coloneqq \\
&\sigma\Bigg(\sum_{i=1}^{n_t} \min_{k,l \in \{-1,0,1\}} \Big\{ d^c\left(day_i^{(1)}, day_{i+k}^{(2)}\right) + \frac{|k|}{2} + \\
&\qquad\qquad d^c\left(day_{i+l}^{(1)}, day_i^{(2)}\right) + \frac{|l|}{2} \Big\} \Bigg)
\end{aligned}
$$

where $\sigma(x) \coloneqq 1/(1+e^{-x})$. The distance measure for cells d^c uses a combination of the Euclidian distance d^c and the adjusted distance that was used in the clustering of the cells. Negative values bring the distance d between the days closer to zero, whereas positive values bring it closer to one.

In d^c negative values are desired if the cells are the same or at least very close. If there is no overlap (low probability of the person being at the same location but being connected to two different cells), the penalty should reflect the distance between masts: As larger differences in a specific time slot make it less likely that the difference is due to a slight deviation from a normal pattern, larger distances should be penalised stronger than small distances. All of the above resulted in the following definition for d^c:

Distance between cells

$$
d^c(c_1, c_2) \coloneqq \begin{cases} -1 & \text{mutual overlap} \\ -0.5 & \text{one sided overlap} \\ \text{NA} & \text{one of the cells is NA} \\ log_{100} d^e(c_1, c_2) & \text{otherwise} \end{cases}
$$

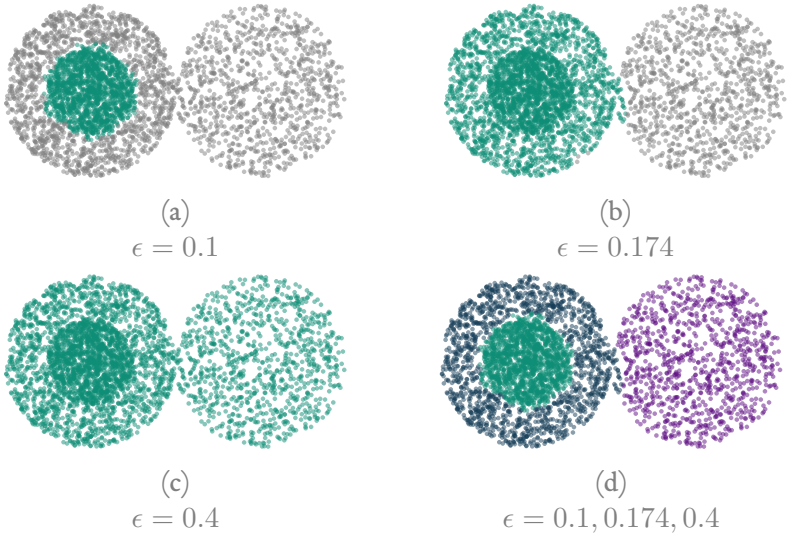
The minimum function in d treats an NA output from d^c as plus infinity. If one of the two parts in the minimum cannot be brought to a real value (i.e. all timeslots in a 1-neighbourhood are missing values), the term is set to zero. Overlap happens if the second cell in question has a Euclidian distance to the first cell that is less than what could be expected based on the conclusions drawn from Figure 3.11. This formulation of the distance is very much related to localised dynamic time warping (Berndt and Clifford 1994) in that the looked for solution is a least-cost path through pairs of cells. The difference however lies in the way this formulation lets us treat missing values. If a reasonable (i.e. close to 0) cost to a connection to a missing value is assumed directly in d^c then timeslots with far away cells are avoided in favour of empty cells in the matching process, making the days seem more similar than they should. Therefore, connections to missing time slots are only allowed when there is no available observation in the whole 1-neighbourhood. The proposed distance measure for days is both positive (because of the sigmoid and the finite values for its inputs) and symmetric (as the days are exchangeable), which are the requirements for DBSCAN.

Clustering days

In the clustering process of the observed user days generally a small ϵ environment would be desirable, resulting in clusters containing only days that have matching cells in many time slots. However, certain clusters simply are not discernible at too low thresholds. Choosing the threshold too large on the other hand creates the risk of not distinguishing between different clusters or clustering days that do not at all represent similar days. To overcome this issue, DBSCAN is applied iteratively with a sequence of increasing ϵ. Days belonging to identified clusters are removed from the set of days to consider for subsequent values of ϵ.

The reasoning behind this iterative process is illustrated in Figure 3.13. A story behind such clusters could sound somewhat like this: There are normal work days with lots of calls and normal work days (also with lots of calls) that are somewhat different, e.g. by a visit to a bar after work. In addition there are weekend days with fewer calls, resulting in less dense days (since no CDR's mean no matches and thus larger distances between the days). Choosing any fixed ϵ can at most separate one cluster from the other two and therefore is not sufficient. Sequential clustering with increasing ϵ first finds the dense clusters followed by the later, less dense ones.

The reconstruction of days (Step 7) is then chosen in a straightforward manner: Given the observations of a day the reconstructor looks for the cluster that is the closest (again using d) to the observation. If there are multiple candidates, take the one with the lowest cluster number is taken, corresponding to the cluster with the smallest epsilon environment and hence the most solid cluster. From that cluster the mode of cells at every time slot is taken, removing those time slots where the mode appears only once (typically early in the morning or late at night). That information is then used to fill in the missing values of the observation. Any time slots that are still missing a value are then filled by the closest non-missing value.



(a) $\epsilon = 0.1$ (b) $\epsilon = 0.174$ (c) $\epsilon = 0.4$ (d) $\epsilon = 0.1, 0.174, 0.4$

Figure 3.13:
Clustering three clusters in the following setup: Two clusters have the same centre but one is a bit more dispersed; the third is further away and even more dispersed. Unclustered points are shown in grey and every identified cluster has its own colour. Choosing a fixed ϵ cannot separate the three and will not find more than one of the clusters. Sequential clustering with increasing ϵ can separate all three.

Evaluation

The two proposed methods were compared with two benchmarks: The first one (denoted "mode by slot") assigns the most frequently seen cell by time slot to the time slots with no observation (i.e. one cluster over all observed days). The second one (denoted mode by time and Weekday/-end) assigns the mode of the cells observed by time slot and an indicator function for Weekends (Saturdays and Sunday) to missing observations (i.e. clusters follow days of the week) and was implicitly or explicitly assumed in (Ahas et al. 2015; Jiang, Ferreira and Gonzalez 2012; Kung et al. 2014; Ranjan et al. 2012).

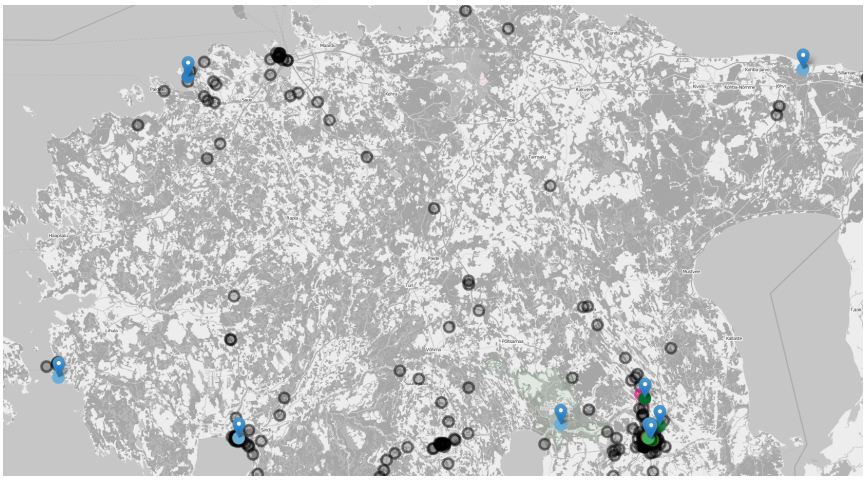
For the evaluation and comparison of the different methods to identify idiosyncratic daily behaviours the distances between predictions and the actually recorded positions were calculated and compared. The predicted location remains constant for every predicted time slot and thus is both spatially (cell size) as well as temporally imprecise. As the GPS measurements sometimes come at irregular intervals all measurements are weighted by the durations of the intervals during which the GPS-position was not updated.

To put the obtained results into perspective also the distances obtained by using the handover data in temporal segments that reflect the actual connection (i.e. not matched to time slots) were calculated. This sets a natural upper limit to the accuracy of the predictions. As clustered cells were used, it is possible that the centroid of the cluster was closer to e.g. the home of a user than any of the individual masts, so it could happen that the prediction had a lower average distance than the cell tower "ground truth".

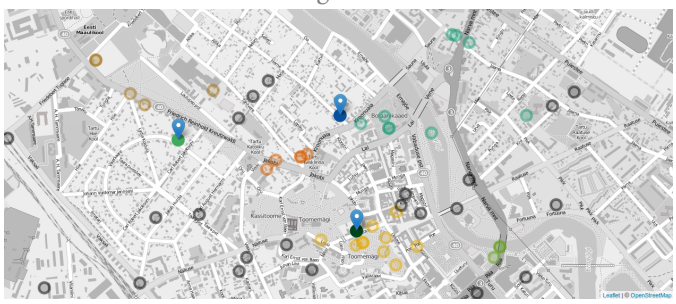
3.4 Results

3.4.1 Real world data

First, an example of the clustering of important locations as used in DAMOCLES is shown in Figure 3.14. Not all significant GPS stops are plotted to avoid overcrowding.

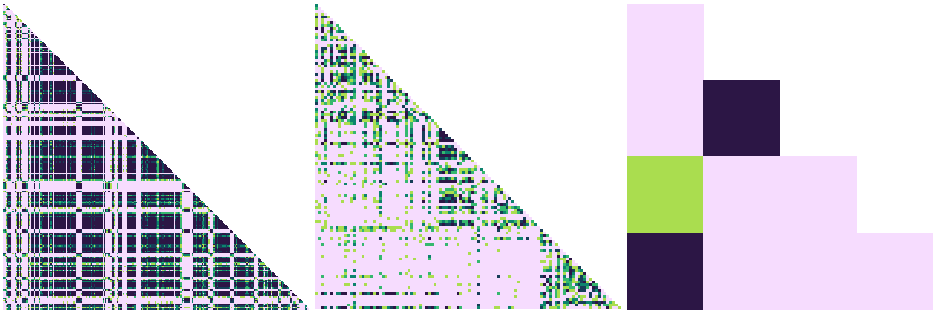


(a)
Large scale



(b)
Small scale

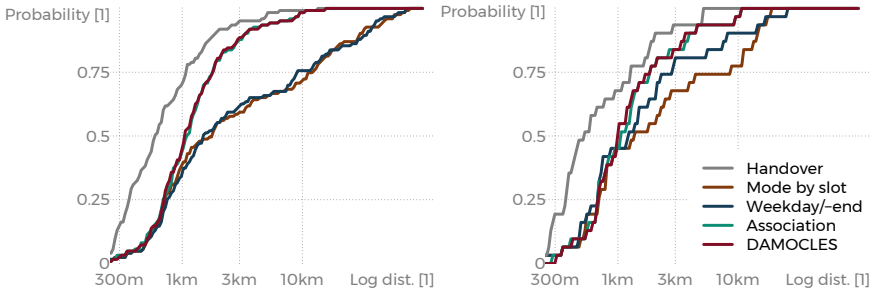
Figure 3.14:
Examples of the mast clustering for a user. Unclustered masts are shown in gray, clustered masts are in color. Markers point to locations that were found important from the GPS signal. Note the different spatial extent of the clusters in the small scale image that is a result of the adjusted distances.



(a) User 187 (b) User 116 (c) User 251

Figure 3.15:
Distance matrices between the days for two users, with the distance between days i and j in row i and column j. The brighter the colour the greater the distance between two days. On the left hand side, two alternating regimes are very clearly visible. In the middle, there are three regimes that follow one another and are separate from each other and on the right hand side, a user with very few data of sufficient CDR, preventing any reasonable clustering.

Distance matrices between user days are shown next in Figure 3.15 for different users exhibiting different behaviour. This is mainly a sanity check, as DAMOCLES clusters based on those distance and therefore certain patterns should be observable in those distance matrices.



(a) All users (b) Users with regular behaviour

Figure 3.16:
Empirical cumulative distribution functions CDF of the median daily distance between the actual position and the reconstruction. On the left hand side the CDF is over all users whereas on the right hand side, only those users who show a regular behaviour are considered, explaining the more clearly visible steps in the function. While the abscissae are without unit, the labels of the ticks have been translated back to be more human readable.

Figure 3.16 shows for all users the CDF of the median of the daily average Euclidian differences for the different methods by user. This way every user gets equal weight, irrespective of the number of days they were under study. The general appearance of the image is the same for all three tested partitions of the day, so only the one corresponding to 12 partitions is shown.

Of course the median cannot convey the behaviour of the prediction in all detail. To shed some additional light also the behaviour for quantiles other than the median in Figure 3.17 is shown, but limited to the DAMOCLES method to avoid overcrowding the image. The other methods show fans of similar width around their respective median curves (not shown).

The sample size is limited, so dividing the population into sub-populations (such as frequent and infrequent callers) leads to results that strongly vary with the individuals, so not make many statements about sub-populations can be made. One that stands out however is the one about users that show a particularly regular user behaviour, shown in Figure 3.16, as they approximately follow the rules assumed by the benchmarks.

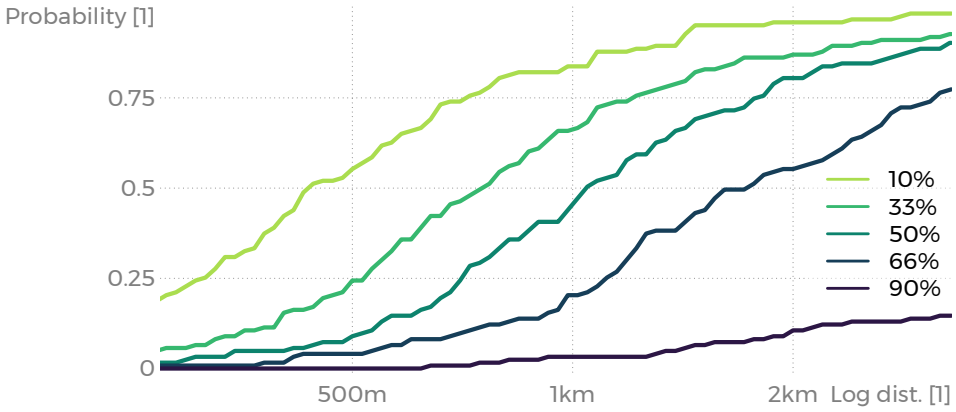


Figure 3.17:
Cumulative distribution function of different quantiles of average daily distances by user for the DAMOCLES method. While the abscissa is without unit, the labels of the ticks have been translated back to be more human readable.

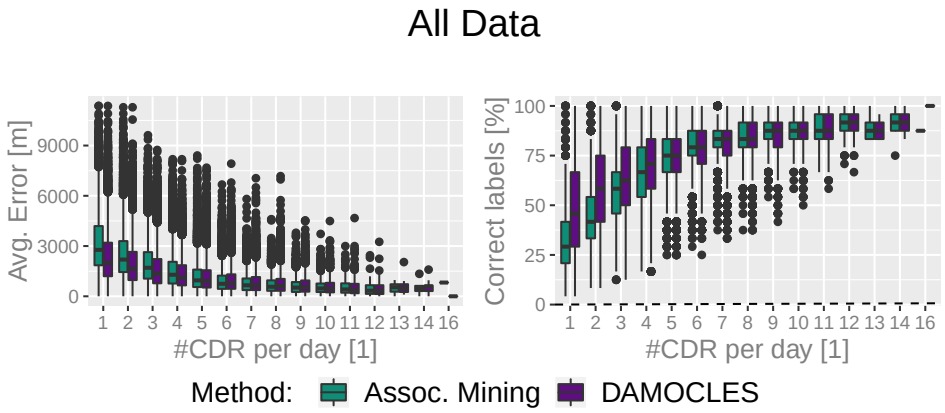


Figure 3.18:
Boxplots with the results for the simulation study. Left: the reconstruction error as daily averages of distances between simulated and reconstructed locations. Right: Daily averages of correctly attributed mast IDs.

3.4.2 Simulation

The results of the simulation study can be found in Figure 3.18, where the overall averages are shown and in Figure 3.19, whereas the results from the simulated users with very few CDR's per day are shown.

Lastly Figure 3.20 shows some reconstructions of daily movements. The reconstruction is plotted in green against the GPS in red and the handover in black.

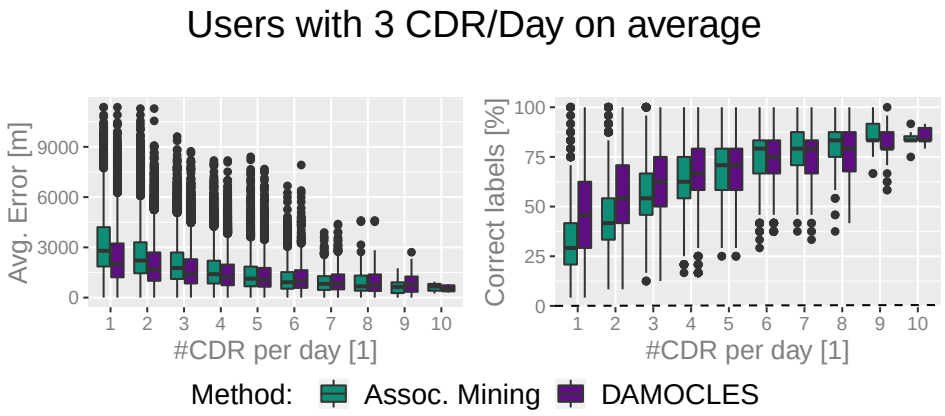
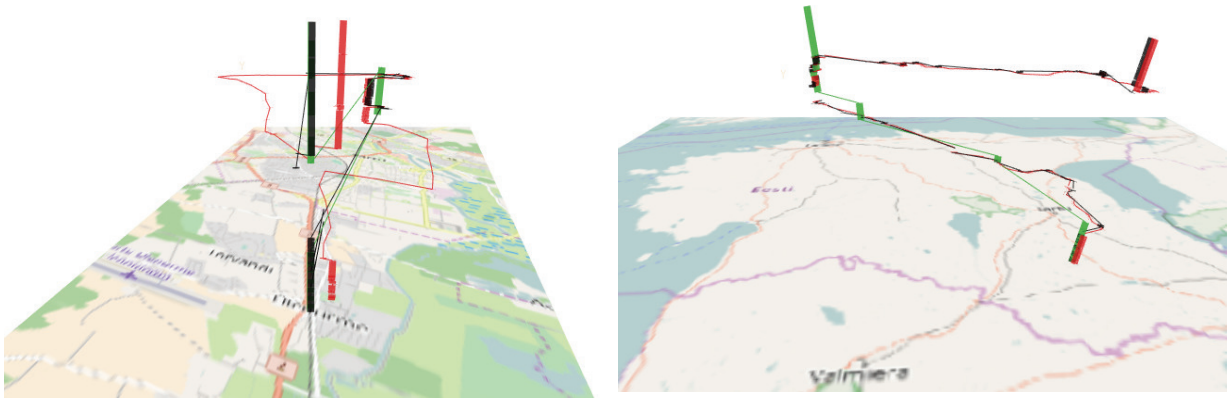


Figure 3.19:
Boxplots with the results for the simulation study for users with 3 CDR's a day on average. Left: the reconstruction error as daily averages of distances between simulated and reconstructed locations. Right: Daily averages of correctly attributed mast IDs.



(a) User 37

(b) User 58

Figure 3.20: GPS (red), handover (black) and reconstructed (green) trajectories of one day each of two users. Clearly the reconstruction can at most capture what is in the handover data and only if there is enough repetition.

3.5 Discussion

Identifying cell clusters that are far apart, as in the large scale panel of Figure 3.14 is easy. It is within cities, as shown in the small scale panel where the tuning of the parameters poses becomes important. If the clusters are chosen too small, identifying similar days becomes more difficult, as the clusters of the same GPS point do not have to bear the same label. Choosing them too big will yield trivial results for people whose important locations are relatively close together.

Clustering

As Figure 3.15 demonstrated, the clustering of days in DAMOCLES is capable to distinguish patterns that can be matched to those in the visualisations of the GPS signal. This is a necessary precondition for the other steps to work. One obvious limitation is that if there are only few days with the required minimum of CDR's, there is no distance matrix to speak of, resulting in the inability of DAMOCLES to do much.

Improvement over benchmarks

As Figure 3.16 demonstrates, clearly the two methods that were used in this case study are better at reconstructing the actual movement of the users than the benchmark solutions, indicating that the patterns captured by them are more helpful for estimating the users' whereabouts. Note that even if the handover ground truth is used, there are days where the average distance is considerable, hinting at an irreducible uncertainty that comes with the nature of cell ID data. This uncertainty can originate from time spent in regions where the cells were large or from incorrectly geo-referenced cells in OpenCellID.

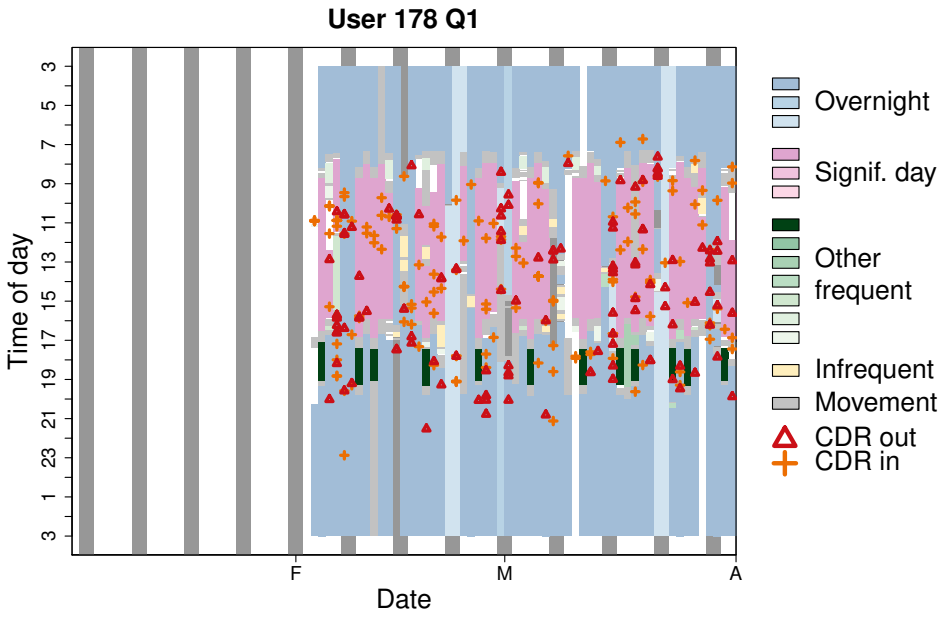


Figure 3.21:
Example of a user with a frequent location (in dark, saturated green, between roughly 17:00 and 19:00) where there is never any CDR activity.

On the subset of users that actually had regular daily and weekly structures, following approximately the assumptions underlying the benchmark reconstructions, the results for those benchmarks can be assumed to be better than average. This can indeed be observed in Figure 3.16. However, still the two more involved methods clearly outperform the benchmarks.

Difficulties and errors

Figure 3.17 shows that for almost all the users there seem to be at least 10% of days that are very poorly predicted. This statement holds true irrespective of method and number of time slots (not shown). The exact number is of minor importance, as the users under study are not representative of any general population and the sample size is rather limited. These days can correspond for example to larger trips without CDR that may start or end at home and thus can be falsely attributed to a prototype day leading to grossly wrong predictions. Alternatively they can correspond to excursions during which no CDR was recorded. They typically combine seeing previously unvisited locations (thus rendering useless any reconstruction based on the past) with stays in rather far away places, resulting in particularly large errors.

Other causes for bad results can be a high proportion of movement throughout a day, such that the restriction of the prediction to the time slots of fixed width prohibits an adequate representation of the movement. Lastly they can be the result of locations that were frequently visited, but never recorded by the CDR. An example of the latter can be seen in Figure 3.21 where there is a clearly discernible frequent location that is visited after what can be presumed to be work, but where there is no CDR that would allow capturing this behaviour in the first quarter of 2015 (and only very few in the rest of the year).

Note that both methods that do not make a priori assumptions are at least as good as the benchmarks even if the users happen to actually work on exactly the days that the benchmarks assume. Neither the association mining approach nor DAMOCLES is visibly stronger for all users on the data collected by YouSense.

Simulation

As for the simulation results, interestingly DAMOCLES seems to work better than the association mining approach on days where there are just very few CDR, as illustrated in Figure 3.18. Most strikingly this is the case for label correctness, where the median lies at nearly 50%, whereas association mining has a median of just over 30%. While less pronounced, similar observations can be made of the reconstruction error, measured by average distance between simulation and reconstruction, where DAMOCLES is less error prone on days with low CDR counts. In both cases the methods start to look similar as soon as 5-6 CDR's are recorded on a day. This also happens to be the approximate value of the threshold needed to reach the saturation point, at which the error reaches the (irreducible) error incurred by the randomly fluctuating starting times of the activities.

The results of Figure 3.18 are qualitatively similar if the total simulation is subset population into the classes identified by the choices for the parameters. Specifically, it is interesting to compare the overall picture with the one obtained from the subpopulation that had 3 CDR's a day on average that you can see as Figure 3.19. Both approaches still can yield reasonable results for users with CDR's in as few as 3 time slots a day on average.

Figure 3.20a shows nice home–work–home pattern that is well captured by the reconstruction, even if the precise timing of coming back home is somewhat off. The distance between the red on the one hand and the black and green trajectories on the other hand shows that the true (GPS) location differs from the GSM location due to the spatial granularity of the mast locations. The figure also shows that the non-GPS signal usually do not capture the exact route that was chosen, as information between stops is often times missing.

In Figure 3.20b, an unusual day for user 58 is depicted. In particular, there was no matching pattern that included the location visited that day. As a consequence, the missing time slots are assumed to be the last known location, which in this instance differs from the information from the GPS trajectory quite substantially.

3.6 Conclusion

In this case study it could be demonstrated that association mining and DAMOCLES can both be used to reconstruct daily whereabouts of users, given their CDR's for an extended period of time. Both methods can capture different habits of movement in ways that do not require a priori assumptions on working days (Jiang, Ferreira and Gonzalez 2012) or working hours (Ranjan et al. 2012).

Association mining and DAMOCLES

Association mining is computationally fast and yields stable results, as shown in the study on the data collected via the YouSense application. However this method is not able to capture the spatio-temporal information underlying the data, and specifically does not distinguish between a small or large spatial error. Also, all the rules learned concern only individual time slots and the big picture of what a typical day as a whole looks like is missing.

DAMOCLES on the other hand is able to find examples of typical days in a way that considers the spatio-temporal characteristics of the underlying data in its decisions. In addition, whole days are considered, which allows for a more interpretable result as well as a superior reconstruction performance on days with only few observed CDR's, as the simulation study demonstrates.

The absence of a priori assumptions means that both tested methods yield their results irrespective of working hours or the days of the week that the movement habits follow. This is a clear advantage, as demonstrated by the worse performance of the benchmarks representing those assumptions in the YouSense study. This can be seen as an indicator that these methods can be used for studying large fractions of a population, where systematic errors on night or weekend workers may bias the findings.

Limitations of DAMOCLES

Apart from its benefits, DAMOCLES also has its limitations. Due to the clustering, it can only work if there are enough days in which there are enough (and dispersed enough) CDR's that allow the distance function to get low enough for clustering. This means that the method does not work for users with constantly very low numbers of CDR's. However, as the "high enough" number only are needed to identify the clusters, the average number of CDR per day can be much lower than for methods that directly reconstruct movement from CDR data (Schulz, Bothe and Korner 2012; Widhalm et al. 2015). Specifically, the simulation study shows that reasonable results can be obtained for users with as few as three CDR's per day on average. Another limitation that is inherent in CDR data is that they can only capture locations where CDR's occur and hence any unreported locations will be missed.

Extensions

There are several ways in which DAMOCLES could be extended. As it is presented here, the temporal regularity of the typical days is not used in order to reduce the assumptions made to a minimum. If one is willing to assume that there is some regularity in the temporal sequence of daily regimes one could easily extend both the clustering and the matching parts of the algorithm to include information on e.g. CDR's from preceding and succeeding days or the day of the week to incorporate ideas from sequential analysis such as for example Rinzivillo et al. (2014) have done. A second extension that could benefit both DAMOCLES and the association mining approach concerns the first and last location on a day. Some users hardly ever have CDR's in the GPS stop segment that covers midnight and therefore both approaches at times fail to detect the first and last stop segments of a day. One way of dealing with this issue could be to include any of the methods from the literature to find sleeping locations (e.g. (Ahas et al. 2010)) and select the first and last locations based on the estimated probabilities of the identified locations.

Lastly, one could develop an integrated approach that combines methods for different amounts of information to reconstruct every day as well as possible. For time spans with high CDR counts one could go for a method as fine grained as (Widhalm et al. 2015), whereas for intervals with fewer observed CDR's, one could use e.g. DAMOCLES. To extract the intervals on which to use the more refined method, a sensitivity study on that method that detects when it breaks down would be necessary.

Models with few a priori assumptions about human mobility are needed when large parts of the population are analysed. Especially minority populations that do not conform to standard assumptions about everyday habits may otherwise be misrepresented. This case study contributed one such method.

Cell ID's remain problematic

Despite all efforts, one limiting factor of approaches in this direction will always remain the imprecision incurred by using cell ID's as proxy for location. Especially in urban areas relatively small differences in distance can make a large difference when trying to infer the semantics of movement. Even the denser masts will not be able to remedy this, as the increased density does not translate directly into more precision, as shown by Figure 3.11. Therefore, while cell based studies can reveal approximate geometries of movement, they will not suffice to extract more detailed semantic information based on that geometry.

Chapter 4

Mode detection from CSD like data

Ye shall know them by their [deeds].
— Matthew 7:16

This chapter is based on a research article submitted to the journal *Computers, Environment and Urban Systems* under the title 'Minimal requirements on spatial accuracy and spmaling rate for transport mode detection in view of an imminent shift to passive signalling data' and is currently under review.
The scientific contributions of other researchers to this chapter were the following: Robert Weibel, Henrik Becker and Kay W. Axhausen all took part in the discussions and decisions during development and helped prepare the manuscript for submission. Kay W. Axhausen ensured access to the data and Henrik Becker contributed substantially to the description of the data in the manuscript. Robert Weibel provided supervision and feedback throughout the project.

4.1 Study setup

As seen in Chapter 3 that even without making assumptions about human mobility other than its repetitive nature, statements about the *habitual* movements can be made. However, ideally one would like to extract more detailed information about the underlying movement, even for special events (Nilbe, Ahas and Slim 2014) or for tourists who don't have an extensive CDR history where they visit (Ahas et al. 2008a). Ideally this includes not only a sequence of all the visited places, but also when and how the people got from one place to the next (Shen and Stopher 2014).

In transportation science questions like the ones above have been the focus of interest for years and have typically been answered with the help of questionnaires or GNSS trackers that have yielded satisfactory results, as described in Section 2.2.

Limitations of cell ID's

However , as elaborated in Chapter 3, the prospects for attaining this goal also by just using CDR are diminished by the cell ID accuracy of CDR. Further, the presented methods matched days to frequent daily patterns and thus are not ideal for days when a person deviates from those habits. Those deviations however, can be the focus of interest to monitor e.g. leisure travel.

In this situation, the case study that is presented in this chapter aims at finding the point between GNSS data and CDR data at which the methods that have been proven to work on GNSS data still can be applied. The term *between* above is of course not properly defined.

Need for trajectories

In order to define it more accurately, the following observations are needed. In order to allow for uninterrupted statements about the movement, gaps of hours, as they are common in CDR's, are not acceptable and so one requirement is that the data can be used as a (simplified) trajectory in the sense that at least the distance travelled from Table 3.2 are at least moderately close to those obtained through a GNSS. The proximity of this sparse trajectory to the true movement is determined by the accuracy of the positioning and by the time intervals at which a position estimate is performed.

Contributions

These two crucial variables – spatial accuracy and temporal granularity – are the ones under investigation in this case study. Starting from a GNSS signal, which is used as the reference, the fixes obtained were subsampled and distorted gradually while the effects of this distortion on the results of transportation mode detection was observed. The contributions of this case study are thus:

- An analysis of how popular GPS-based transportation mode detection algorithms work under deteriorating conditions in terms of both spatial accuracy and temporal granularity of the underlying data.
- Recommendations derived from that analysis on the necessary data quality for purely passive tracking of mobile phone users.
- An assessment of different transportation mode detection techniques on data that has breadth in the user base but less than perfect labelling.

4.2 Materials and methods

The aim of this case study is to demonstrate how close to a GPS signal passively sensed data needs to be in order to allow transport mode detection using the techniques typically associated with GPS data. For this, the GPS data are progressively distorted spatially and subsampled more granularly temporally in order to find the limits at which the traditional GPS approaches to transport mode detection are no longer useful, thus denoting the minimal requirements on accuracy and frequency of passive tracking.

The necessary terminology for this case study and the related work that explains some of the choices made here can be found in Section 2.2.

4.2.1 Data

The data was collected as part of a pilot study for a new *Mobility as a Service* (MaaS) offering of the Swiss federal railway company SBB [1] with 138 participants across Switzerland. They were selected to cover a variety of living conditions so that the experiment covered the whole country, and different types of spatial backgrounds (from small village dwellers to residents of large cities). The MaaS offering included both unlimited use of public transportation within Switzerland as well as the lease of an electric car (in addition to the cars and bicycles already available in their household). Hence, respondents showed a diverse and highly multi-modal travel behaviour.

As part of the study, each participant had to record a travel diary for the whole duration of the study (about one year) using a re-branded version of the MotionTag smartphone-app.[2] The app uses the device's location services to record coordinates, transmits them to a server, where the records are classified into trips and activities using proprietary algorithms. Respondents were able to review their records from within the app and were asked to confirm or edit activity types and modes as well as to report any erroneous records. The sampling frequency was set to 1Hz, but was of course heterogeneous due to e.g. signal loss that happened regularly in for example trains or tunnels.

Data used

A partial dataset containing all records and trips made between March and August 2017 was available for this research. The data contains both the raw data (*waypoints*) and the annotated trip data (*tracks*):

- *Waypoints*: 21119962 observations (after subsampling to the highest temporal granularity used here, removing the most obviously wrongly labelled trips and using only trips that remained within Switzerland) containing user ID, timestamp, longitude, latitude;

[1] https://www.sbb.ch/en/travelcards-and-tickets/railpasses/greenclass.html

[2] https://motion-tag.com

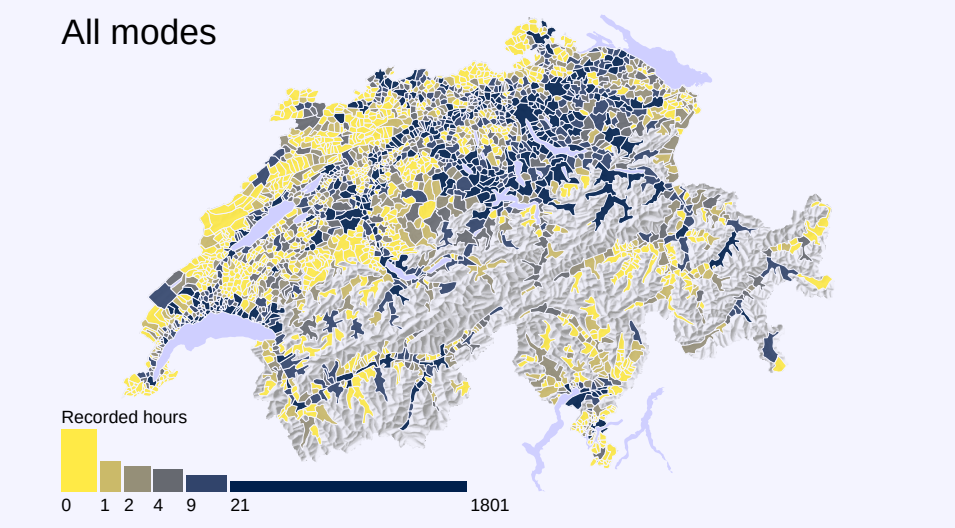


Figure 4.1:
Distribution of the data points in space. The municipalities are colour-coded by how many hours worth of data was collected in them, with the colour code representing the quantiles of municipalities. There were much more recorded hours in larger cities, especially Zurich, which reflects the distribution of the population.

- *Stages*: 117 091 observations containing user ID, start date/timestamp, end date/timestamp, distance, PostGIS geometry, detected mode[3], confirmed mode, user comment. 96.1% of the tracks have a user-confirmed transport mode.

The geographical distribution of the data points is shown in Figure 4.1. The municipalities with most recorded data lie along the main traffic routes of the country, whereas the more rural areas feature fewer hours of recordings. The collection covers the entire country with the exception of the Alps.

Areas that are not considered habitable by the Swiss Federal Office of Statistics [4] (lakes, glaciers and rock) are not considered and left blue (lakes) or in relief (glaciers and rock).

The user-confirmed trips from the *tracks* dataset represents the ground truth, which needs to be replicated by the *waypoints* dataset as raw data.

Table 4.1:
Most common label sequences on trips. Even among the most common sequences there are some that are not what one would expect from theory (e.g. Train without walking stage leading up to it).

Mode Sequence	Count
Car	38 382
Walk	22 251
Car, Walk	3 812
Walk, Car	2 909
Bicycle	2 842
Train	2 098
Train, Walk	1 439
Walk, Train, Walk	1 025
Walk, Train	987
Bus	699
Total	90 515

[3] The options were: airplane, boat, coach, bus, tram, train, car, bike, ski and walk. Trips containing the extremely rare modes of ski, coach, airplane and boat were removed for the purpose of this study

[4] https://www.bfs.admin.ch/bfs/en/home.html

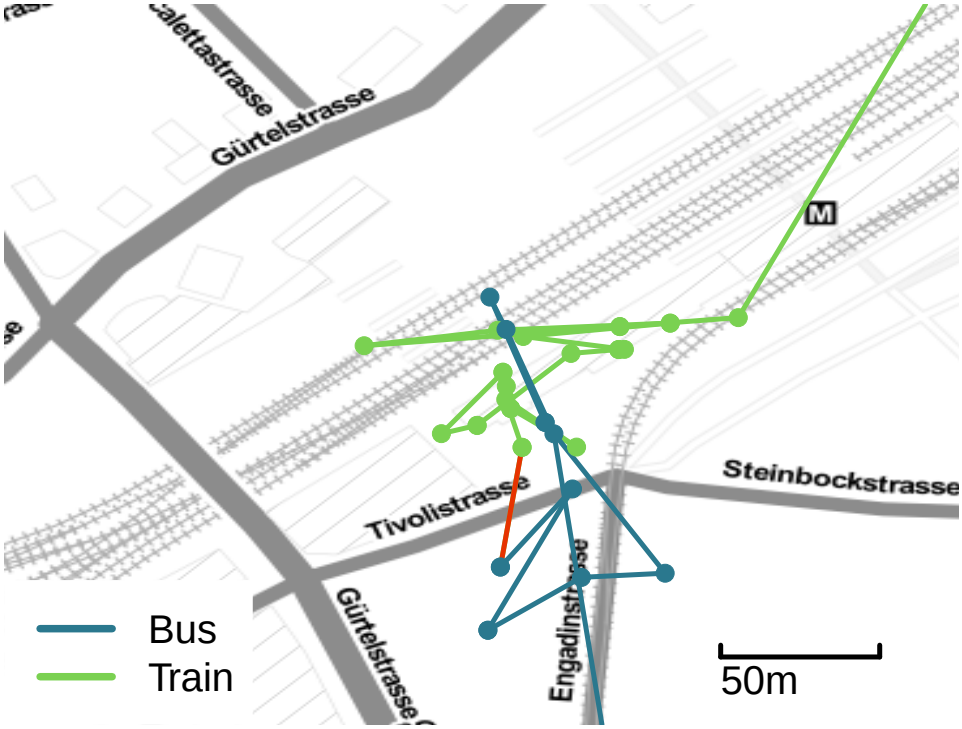


Figure 4.2:
Detail of an annotated trajectory near the main station in the city of Chur. The red line highlights two consecutive points with different non-walk modes of transportation without a walk stage between them. Here the situation of the GPS fixes would suggest a rather stationary behaviour indicative of a walk stage.

Labelling

For setting the ground truth labels, the app only allowed participants to delete or report erroneous records, but not to modify the path or the segmentation. This gets reflected in the most common sequence labels in the dataset shown in Table 4.1: For example the sequence "Train" appears about twice as often as the sequence "Walk, Train, Walk" indicating that some segments may have not been identified correctly. In particular, this affects access/egress walk stages, which occur substantially less often than expected.

One example of this can be seen in Figure 4.2, where the last train point is immediately followed one labelled *Bus*, even though there clearly seems to be a time where the user in question was moving very slowly in the train station area and semantically the labels as they are make little sense.

In Table 4.1 the unweighted counts are listed. It has to be noted that for different modes the typical duration of a stage of those modes may vary. To provide the full picture, also the number of points (on the 5 second granularity) are provided in Table 4.2.

As illustrated in Figure 4.3, GPS signal loss is most common on trains, so the number of hours travelled in trains was higher than suggested by the point counts. It is evident that the data are heavily skewed in favour of the *Train* and *Car* modes, which has to be accounted for when training the classifiers lest one obtains too optimistic results.

Mode	Number of points
Car	6 091 407
Train	3 050 785
Walk	1 984 240
Bike	511 906
Bus	223 977
Tram	116 760

Table 4.2:
Number of ground truth labels (on a 5 second basis) for the modes under study. The distribution is very uneven, but reflects the behaviour of the population under study over the time span of the data collection.

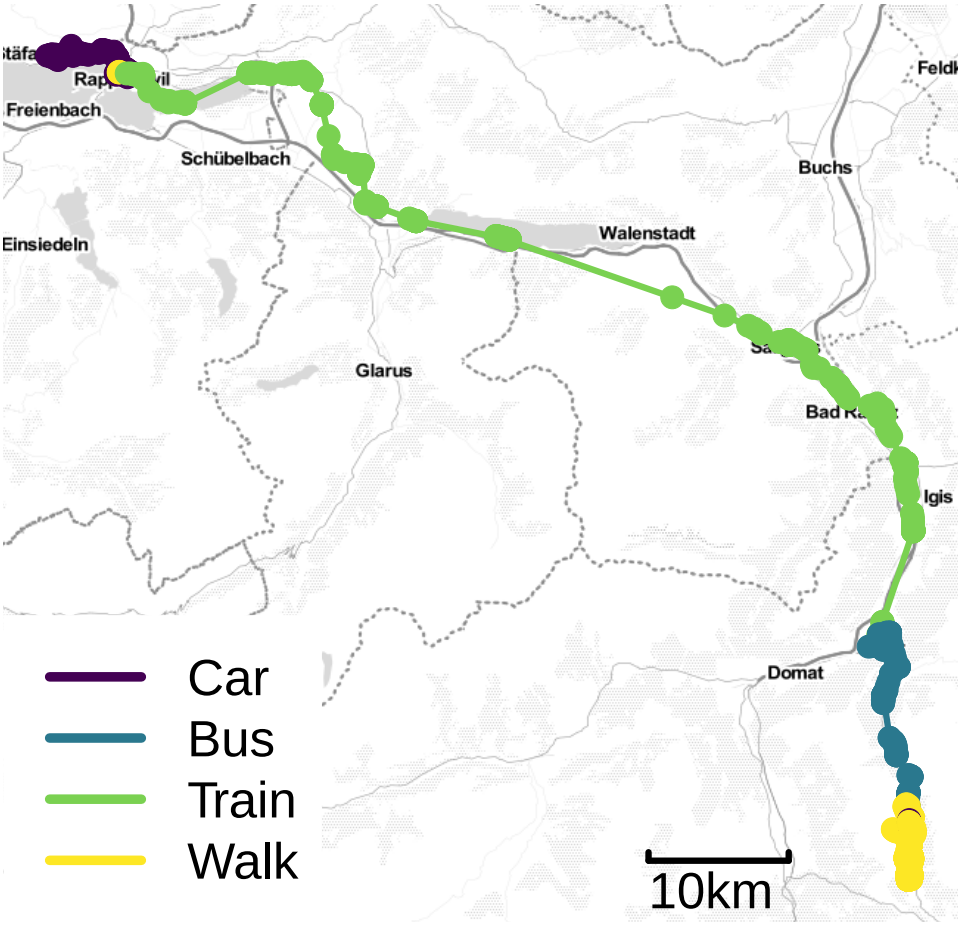


Figure 4.3:
Example trajectory of one user for one day. The train stage of the trajectory is clearly most affected by GPS signal loss. Other modes of transportation seem to suffer less, particularly walking usually has good coverage. Note that even the GPS signal is not perfect, as indicated by the 'shortcut' the signal takes south of Walenstadt.

4.2.2 Subsampling of the data

For the subsampling, a relatively pragmatic approach was used. Time was partitioned into episodes of equal length (e.g. 5 minutes) and for each of those episodes, the first observation was taken. Given that the original sampling rate was about 1 Hz, the unevenness of the time differences between the retained points should not be materially increased by the simplicity of the scheme.

t	Decision	Reason
16:00:00	Keep	First in Interval 16:00-16:05
16:01:00	Drop	Second in Interval 16:00-16:05
16:14:59	Keep	First in Interval 16:10-16:15
16:15:01	Keep	First in Interval 16:15-16:20

Table 4.3:
Illustration of the temporal subsampling. The actual data was collected at 1 Hz. The fictive timestamps given here are simply used to demonstrate the subsampling method.

This regime does not affect the parts of the data where the original signal was already coarser, due to e.g. signal losses in trains and tunnels. The highest temporal resolution was chosen to be 5 seconds which is considerably higher than what seems to be available for passive data today, but may become a reality with ultra dense 5G networks (Koivisto et al. 2017). The lowest temporal resolution was chosen to be 5 minutes, which is in the order of magnitude of individual short trips. Including even lower levels of temporal resolution would lead to problems when identifying the trips as such. Yet, data quality does not appear to be a problem. In recent statements, Swisscom, the largest Swiss mobile phone network provider, reported to collect 20 billion events per day.[5] With 6.6 million mobile phone numbers under management as of 2017[6], this corresponds to a 30 second resolution on average. Given a certain number of actually unused phones, the actual sampling rate may be even higher. Table 4.3 illustrates the subsampling used (assuming 5 minute intervals).

4.2.3 Distorting the data

As spatial distortion jointly normally distributed, uncorrelated pseudo-random errors were added to the position obtained from the GPS signal. The position to be distorted was not latitude and longitude directly, but the x and y coordinates in the EPSG 2056 reference system that correspond to the measured positions obtained through GPS. The standard deviations of the error that were used were 0m (no distortion), 25m, 50m and 100m. The upper bound reflects the order of magnitude to be expected from current LTE trilateration (Müller et al. 2016). The other extreme, 0, reflects the uncertainty inherent in GPS which has been shown to be small enough for traffic mode prediction (cf. Section 2.2). The upcoming positioning data from 5G promise to be even more accurate (Koivisto et al. 2017) than GNSS's, but the effects of that cannot be tested in the setup of this study.

[5] https://ict.swisscom.ch/2015/11/from-big-data-to-smart-data-traffic-optimization-using-mobile-network-traces/

[6] https://www.swisscom.ch/content/dam/swisscom/de/about/investoren/documents/2017/2017-q1-zwischenbericht-de.pdf

4.2.4 Methods

The focus of the work in this case study lies on the effects of the deterioration of the positioning signal and not primarily on the relative merits of one transportation mode detection method compared to another. Therefore, lest the results be consequence of idiosyncrasies of a certain combination of overall classification strategy (pointwise, smoothed pointwise or segmentwise) and classifier, respectively, most relevant cases are covered. As for the features, only those that can nowadays be assumed to be available from passively tracked mobile phone data are used, namely position fixes and geographic information.

4.2.5 Terminology

As already stated, the atomic unit of measurement are the individual *fixes* that denote information (such as the position, but also includes features such as proximity to public transport) at a given point in time.

Segments and Stages

A set of temporally contiguous fixes that are semantically very close can be grouped into a *segment*. However, there are competing notions of a segment, reflecting different ontologies of movement. A set of contiguous fixes that shares a common mode of transport is called a *stage* or an *inferred stage* depending on whether the ground truth or the inferred labels are used as the basis for the grouping.

On the other hand a *segment* is set of contiguous fixes that share characteristics pertaining to the displacement over time implied by the fixes. Those segments are used in some of the used classification approaches to yield more stable or more accurate results. The term "segmentwise classification" always refers to a classification based on that type of segments, as at the time of classification the delineation of the stages are unknown.

Both kinds of division of trips are partitions of the fixes used, i.e. every fix belongs to exactly one segment and exactly one stage. However, the segments and the stages need not coincide. For example, it seems very possible that a bus stage comprises many move segments (one for every move between two consecutive bus stops) whereas it will only be a single stage, as the mode of transport is *Bus* for the whole journey.

Trip

Finally a *trip* denotes the smallest contiguous set of fixes that are deemed a journey between two places where the person performed some meaningful activity (e.g. a trip between home and work). The splitting of the raw data into trips is not part of this work, as the stays were for the most part very clearly discernible. This assumes long stays with short times of movement in between, as was already observed in other studies (Burkhard et al. 2017).

4.2.6 Features

For the "raw signal", the distorted and subsampled data mentioned in Section 4.2.3 are used. To this raw signal, information that would be available to any service using passively sensed positioning data in many countries was added: Speeds that are calculated on consecutive position fixes, distances to public transport facilities relevant to the study area and quantiles thereof over moving windows. The relevant public transport facilities were both the stops (point data) and routes (line data) of buses, trams and trains.

GIS information

The positions of the public transport stops were obtained through the open data portal for public transport (Swiss Federal Office of Transport 2018), whereas the routes were obtained through Open Street Map (OpenStreetMap Contributors 2017). While there is an ongoing ongoing debate about the respective merits of authoritative and volunteered data, the decision made here was to use official data where available, and volunteered data where necessary.

Moving windows

For the moving windows over which the quantiles were to be calculated, 130 seconds was chosen. This was big enough to allow multiple points within the windows for all but the largest temporal granularities and was in line with the orders of magnitude that can be found in the literature (Bolbol et al. 2012; Ellis et al. 2014; Stenneth et al. 2011).

For the quantiles of the featured to be calculated over those moving windows quartiles were taken. The goal was to have a value representing the central tendency and two that represented high and low values respectively. To avoid the detrimental impact of outliers the choice was made against taking the averages and extremal values sometimes found in the literature (Bohte and Maat 2009; Gonzalez et al. 2008).

For all the numeric values (that incidently all happen to be non-negative), a log-transform was added to help less robust classifiers and used scaling to achieve zero mean and unit variance for all of them.

In addition, for all points the quartiles of the above features for all points by the same user that were recorded within a certain radius and a temporal window were added. The reasoning behind this is that a user may move through the same locations using the same mode of transport, in which case having data from the past could contribute to averaging out errors incurred by the imprecise tracking.

4.2.7 Classification methods

As mentioned, the aim was to have representatives of the most common approaches to classifying modes of transport in this study: Pointwise classification, pointwise classification followed by some smoothing, segmentwise classification, and approaches integrating segmentation and classifying in one step.

Pointwise classification

Pointwise classification is straightforward and tries to find a mapping from the features of every individual point in a trip to the most likely mode of transport. The resulting stages are implicit and based on how many consecutive points share the same predicted label. As classifiers KNN, logistic regressions (LL), random forests (RF), and support vector machines (SVM) were used. In addition and not fully compliant with the idea of a pointwise classifier, conditional random fields (CRF) were used. All of those methods have been used with varying degrees of success to classify transport modes and a more in-depth discussion can be found in Section 2.2.5.

For the optional smoothing of the inferred labels, different schemes were used. The first is a simple majority vote over a number of points that would correspond to two minutes if the points were regularly sampled (e.g. 4 for the case where the points are subsampled to 30 seconds). However, there were always at least 3 points, such that even for the temporally coarsest case there would always be real smoothing.

The second smoothing approach uses a Hidden Markov Model (HMM) on the predicted probabilities of the labels, following the ideas in (Nitsche et al. 2014). The HMM was trained in a supervised fashion on the training data using the predicted class probabilities on the training data in combination with the true labels and applying the fitted HMM on the outcome of the predictions for the test data.

Finally, the CRF idea from the pointwise classification was re-used as a post-classification smoother. The training procedure was the same as for the HMM smoother.

Segmentwise classification

For the segmentwise classification every trip was partitioned at spatiotemporal points where the speed was below 1 km/h for 130 seconds. Such simple thresholds on speed and duration are quite common in the literature (Biljecki, Ledoux and Oosterom 2013; Chung and Shalaby 2005; Stopher, FitzGerald and Zhang 2008) and there would conceivably be some accuracy to be gained by devising more sophisticated segmentation schemes.

To the segmentwise classifications two smoothing regimes described above were also added. While the smoothing on the pointwise classification is mainly motivated by the elimination of stray labels, the main reason to also apply it on the labels for the segments is to avoid unlikely combinations such as car-bike-car.

It made sense to also include a method that combines elements of segmentation and classification. Conditional random fields were chosen, as for the smoothing above. As the method encourages realistic sequences of labels, it has to receive special treatment when interpreting the results. While other approaches such as Recurrent Neural Networks also could be useful here, they were not included here, as they are (still) not well established in this domain. Their properties make them interesting candidates for mode detection though and exploring them would make for interesting future work.

4.2.8 Evaluation of classified results

Rooted in the different applications for which transport mode detection is being used, there is a distinct lack of consensus as to how mode of transport detection should be evaluated and there is no benchmark dataset on which all methods are evaluated. While it is not possible to alleviate the second problem here, an attempt to provide different kinds of evaluation metrics that allow accommodating different kinds of questions can be made.

Accuracy

One of the more popular metrics is the accuracy which counts the percentage of correct labels in the evaluation dataset (Ellis et al. 2014; Reddy et al. 2010; Semanjski et al. 2017). This measure is well suited if only the overall proportion of the different kinds of transport modes is important, e.g. in the context of discussions around the (time weighted) mode split. It may, however, be somewhat problematic if the dataset is highly skewed, as the less frequent transport modes will tend to be under-represented in the labels.

Confusion matrix and Cohen's Kappa

An alternative is precision and recall by mode of transport, e.g. reported using a confusion matrix (Mäenpää, Lobov and Martinez Lastra 2017). This has the advantage that if one mode is of particular interest, the error associated with it can be directly read off. However, confusion matrices can in general not be ordered and thus it is not possible to determine a "best" method. To have a single number for comparisons, Cohen's Kappa can be calculated to summarise the matrix (Bolbol et al. 2012; Huss et al. 2014).

Edit Distance

Lastly if one is interested in "representative" trips, then the sequence is of particular interest and measures of differences between sequences must be used, such as the edit distance (Chen, Özsu and Oria 2005). This measure has been extended for information needs that go beyond the sequence as such (Prelipcean, Gidofalvi and Susilo 2016), but as the aim is to present how accurately the methods describe the sequences of transportation modes, this case study is limited to the edit distance.

4.2.9 Cross-validation

Typically, studies limit themselves to splitting the data into training and validation datasets and reporting the point estimates of the chosen evaluation criteria. However, to also gain insights on how strongly those estimates can vary, a 10 fold cross-validation was performed. This cross validation is done on the user level, i.e. splitting the users into 10 bins, 9 of which are used in training for every fold. Thus, producing results that are overly optimistic as a result of data of the same users being used in training and testing are avoided.

4.3 Results

First, the effect of the deterioration of spatial accuracy on the pointwise classifiers will be shown, followed by the results from the segmentwise classifiers (both without longitudinal data). Following that, a list of consequences of deviating from some of the choices made for this study will be given. The section ends with the confusion matrices of the classification.

Results not shown

In this chapter only the results from the K-nearest neighbours (KNN), Random Forests (RF) and Conditional Random Fields (CRF) classifiers are given, to avoid information overload. The first is chosen as a benchmark that serves as a clear lower bound of what one would expect from a classifier. RF was kept as it produces the best results of the simple (i.e. truly pointwise) classifiers while sharing with the others the overall behaviour. CRF as a classifier that inherently considers sequences of points can be expected to behave differently from the rest and therefore warrants discussion.

On the classifiers whose results are not explicitly given here the following brief statement provides an impression of how they performed: The logistic regression overall yielded results that lie somewhere between those obtained from KNN and RF. SVM algorithm used has for fits (Chang and Lin 2011) is above quadratic (Pedregosa et al. 2011), preventing it from being applied to the whole training set. Therefore it was only applied on a subset of the training data. SVM suffered less compared to random forests, as the number of segments in the training set is significantly smaller than the number of points, meaning that the SVM classifier sees a larger proportion of the points. Even in the better case of the segmentwise classification, the results of SVM classification remained below those of the RF.

4.3.1 Pointwise classification

The first results shown are those from the non-distorted data and can be found in Figure 4.4. The numbers from the pointwise classifications are what could have been expected.

Non-trivial classifiers such as RF clearly outperform KNN, because (in the case of RF) they look at more relevant neighbourhoods than simple spheres do. RF in turn gets dominated by CRF, which again is not too surprising, as CRF's can look at more than just the features of a single point to determine its class.

In terms of accuracy and Kappa, for the non-CRF classifiers, sparser temporal sampling seems to coincide with better results. Note that in a sparser sample, features such as velocity average over a longer time, resulting in more context information being available in the features of a single point. After smoothing, however, the results of different temporal granularities are comparable within a single classifier.

The edit distances of both KNN and RF are well above 1 (and at times above 10) and therefore indicate that those methods should not be used, if one is interested in the sequence of the transport modes. CRF's on the other hand have a markedly lower edit distance and the average of edits needed to each sequence is well below one.

When smoothing is added to the picture, some of the advantages of CRF's get evened out, particularly in terms of accuracy and Kappa. Interestingly HMM seems to be particularly bad for KNN and good for RF, whereas the inverse is true if instead the CRF smoother is applied.

Before turning to the results of the spatial distortion it has to be stated that the results are below some found in the literature. Please refer to Section 4.3.4 for a possible explanation.

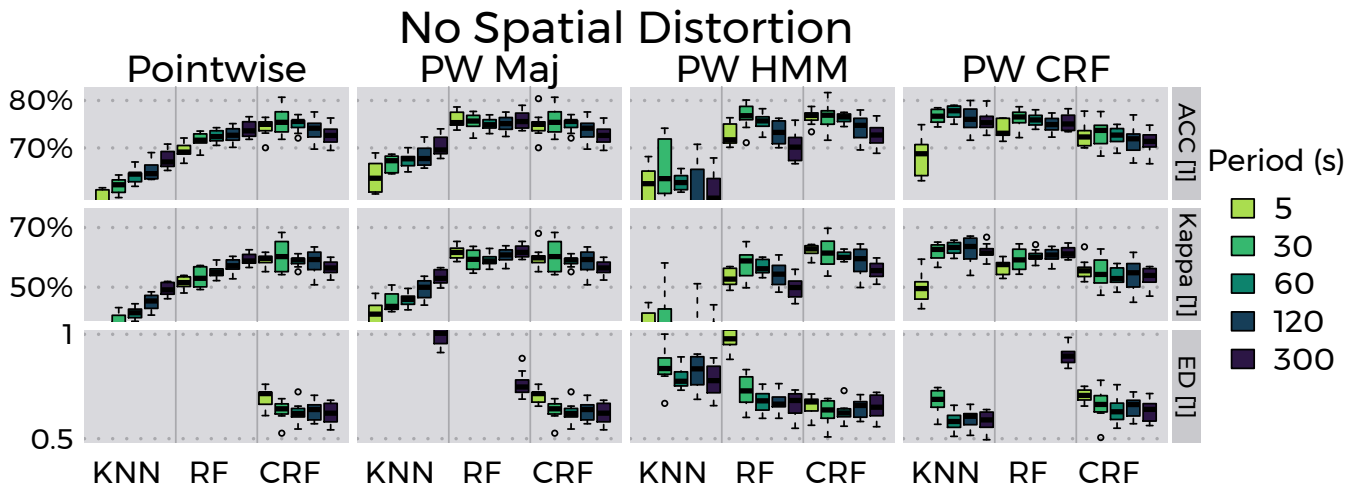


Figure 4.4: Results of the pointwise classification on the non-distorted points. Top row: Accuracy, Middle: Cohen's Kappa, Bottom: Edit Distance. The columns contain (from left to right): Pointwise estimator, smoothed with majority vote, smoothed with HMM, smoothed with CRF. Every plot contains the results grouped by classifier, ordered and coloured by temporal granularity. Every coloured box represents the ten values from the cross-validation.

If one applies a spatial distortion, the pure pointwise classification drops markedly. The effect is stronger, the more temporally fine-grained the data are, as was to be expected. While the effect is somewhat mitigated if the pointwise results are smoothed, 5-second intervals still do not seem to be very useful for direct classification. The combination of a CRF initial classifier and a HMM smoother yields relatively stable results throughout the distortions.

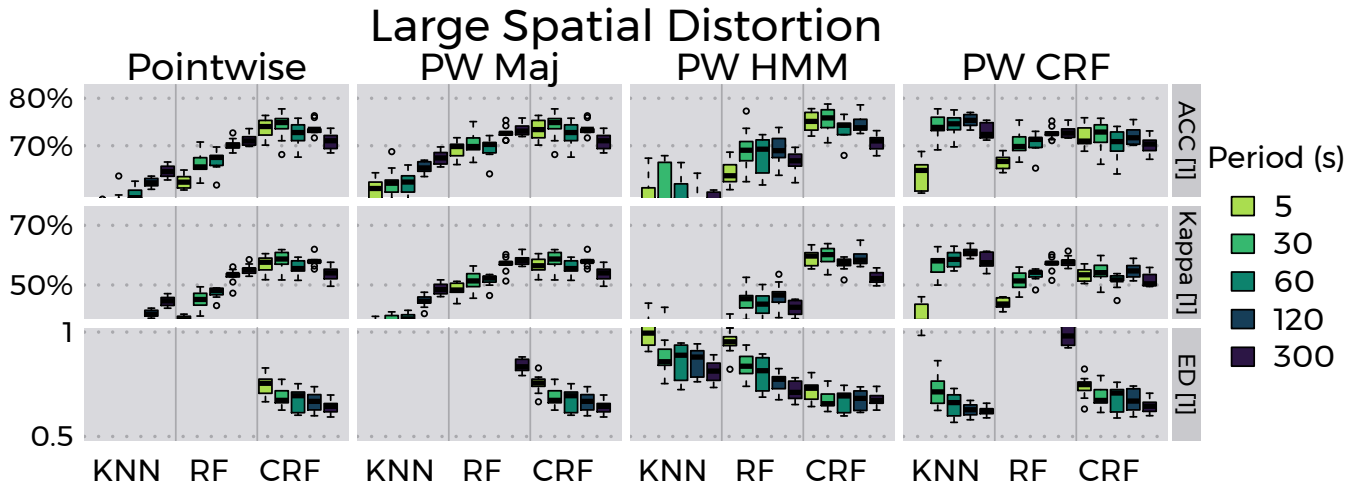


Figure 4.5: Results of the pointwise classification with strongly distorted points.

4.3.2 Segmentwise classification

When looking at the results from the segmentwise classification – recall that the segmentation is neither learned nor known *a priori* but the result of thresholds – there are several striking differences to the pointwise classification.

The two 'simple' classifiers KNN and RF benefit significantly from having features based on move segments, whereas the CRF classifier cannot benefit from them and now has results very similar to the very simple KNN classifier. RF, however, now obtains results that are an improvement over the best of those from the pointwise classification.

The second striking feature of the results is that any smoothing applied to the obtained results does no longer improve them. As a last difference to the pointwise case, the HMM smoother seems to produce distinctly worse results than the other two.

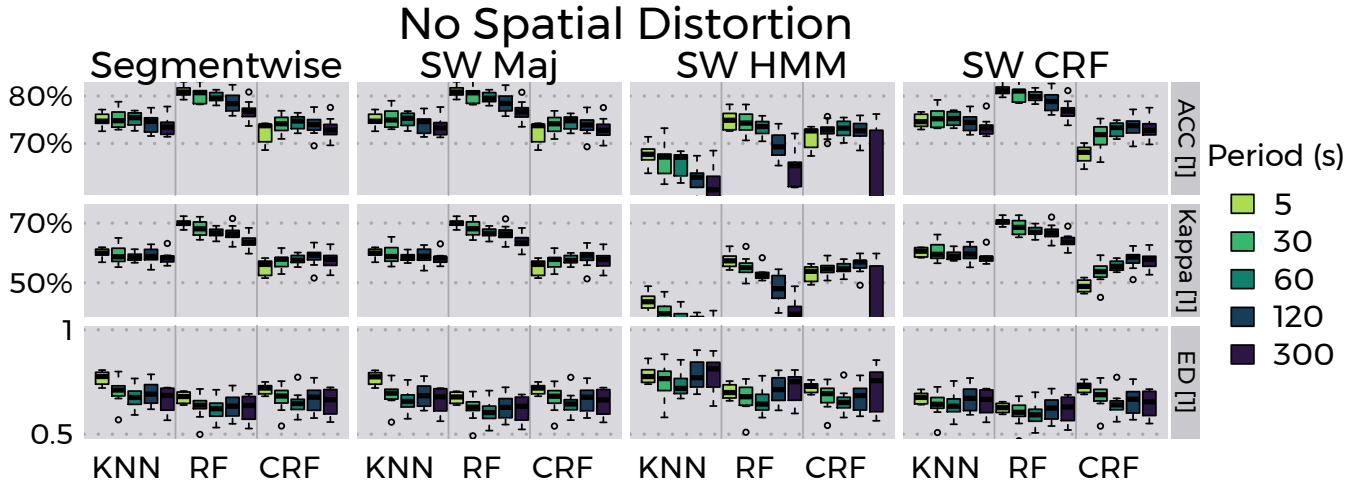


Figure 4.6: Results of the segmentwise classification on undistorted points.

In stark contrast to the pointwise case, classifying segments seems a lot less sensitive to spatial distortions. It can be assumed that this finding echoes the one from the pointwise case, where the results obtained from temporally more coarsely resolved points were better. In both cases, the features are influenced less by distortions of the same order of magnitude, as they are based on points that are further apart. As particularly motorised segments (that abound in the data) can easily be rather long due to a lack of stops, the calculation of overall displacement and median speed are hardly affected by distortions, as they do not accumulate.

When comparing between Figures 4.7 and 4.5, i.e. the pointwise and the segmentwise classification on distorted data, it is striking that even with all the smoothing the results of pointwise classification are simply not as good when classifying in a pointwise fashion than when classifying by segments.

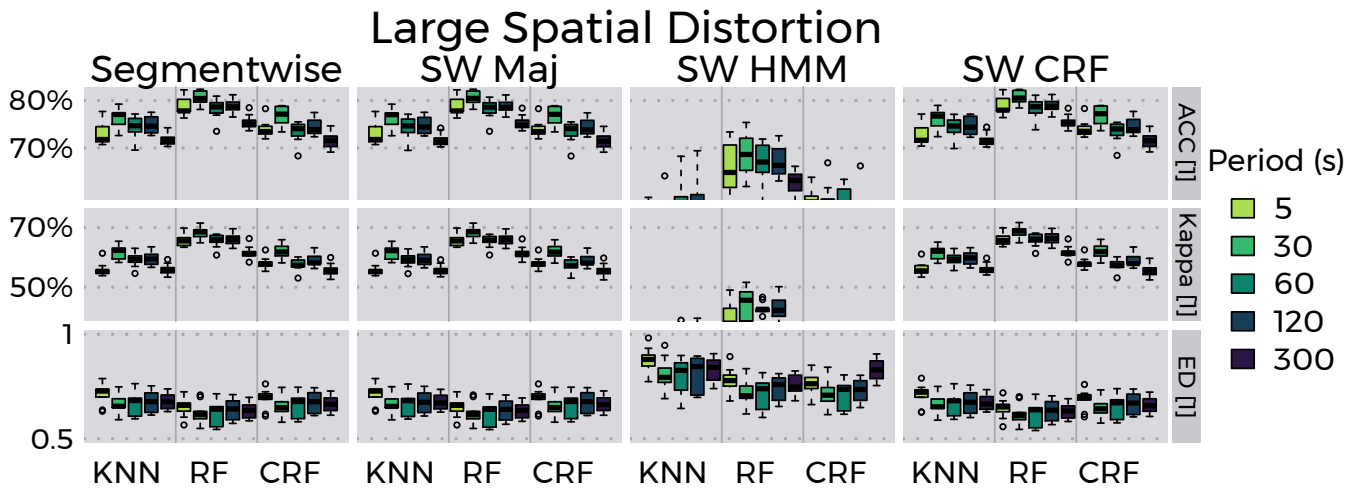


Figure 4.7: Results of the segmentwise classification with strongly distorted points.

4.3.3 Including longitudinal information

Including the longitudinal data does not change the picture dramatically. While there are some positive effects for the pointwise and smoothed classifications, especially for the noisy data the effect is smaller or even reversed in the case of segmentwise classification, where the best results are achieved.

4.3.4 Sensitivities

A complete run through all the possible combinations of choices for all the parameters would not be possible due to the combinatorial explosion of different cases. However, to get some ideas about how certain choices could affect the results shown here a a sensitivity analysis was conducted for a few select parameters that were held constant in the main experiment. The results of these sensitivities can be seen in Table 4.4, where the baseline corresponds to the 30-second and non-smoothed results from the RF classifier in table 4.6. The classifier, smoother, temporal granularity, and spatial uncertainty were held constant for all the comparisons.

Sensitivity	Accuracy	Accuracy SD	Kappa	Kappa SD	Edit Dist	Edit Dist SD
Baseline	80.43	1.47	0.69	0.69	0.62	0.05
No GIS Information	79.57	1.57	0.67	0.67	0.60	0.05
Generous segmentation	77.21	1.16	0.64	0.64	2.03	0.13
CV by trip	80.46	0.35	0.69	0.69	0.61	0.01
GT segmentation	90.03	0.42	0.84	0.84	0.24	0.01

Table 4.4: Median values and standard deviations for the cross-validated quality measures. The temporal granularity was always 30 seconds and the spatial distortion was not present to produce the results. See Section 4.3.4 for an explanation of the sensitivities.

The sensitivities tested were the following:

Table 4.5:
Confusion matrix with no spatial distortion for the combined segmentation and classification problem.

	Car	Bike	Walk	Train	Tram	Bus	Precision
Car	90%	45%	20%	10%	12%	57%	83%
Bike	0%	32%	1%	0%	2%	0%	86%
Walk	3%	12%	57%	2%	19%	10%	77%
Train	7%	11%	21%	87%	23%	20%	76%
Tram	0%	0%	1%	0%	45%	1%	53%
Bus	0%	0%	0%	0%	0%	12%	56%
Recall	90%	32%	57%	87%	45%	12%	

Table 4.6:
Confusion matrix with large spatial distortion for the combined segmentation and classification problem.

	Car	Bike	Walk	Train	Tram	Bus	Precision
Car	89%	42%	20%	7%	14%	50%	83%
Bike	0%	19%	0%	0%	1%	0%	87%
Walk	3%	20%	55%	1%	25%	11%	75%
Train	8%	19%	24%	91%	31%	35%	75%
Tram	0%	0%	1%	0%	29%	0%	57%
Bus	0%	0%	0%	0%	0%	3%	48%
Recall	89%	19%	55%	91%	29%	3%	

- *No GIS Information* removes all GIS dependent features from the data. This should allow for a cost-benefit consideration of taking the effort of adding this type of information when designing a system for passive tracking.
- Generous segmentation corresponds to an alternative choice of segmentation parameters that creates more segments, nameley whenever the speed falls below 10km/h (for any length of time).
- CV by trip does the cross-validation treating every trip as equal instead of cross-validating by user.
- GT segmentation, in addition to cross-validating by trips uses the ground truth segments for classification.

4.3.5 Confusion matrices

Lastly some of the obtained confusion matrices will be shown in this section. Based on the results seen for the different temporal windows, the matrices shown are those at the 30 second temporal granularity and the RF classifier based on segments – both inferred and ground truth.

In Table 4.5 it can be seen that for the undistorted data that the modes that are by far the most common in the dataset (car and train) get classified correctly most of the time, with a recall of about 90% each. This despite the fact that the skewness in the labels was accounted for when training the classifiers. However, the same results cannot be obtained for the modes that are less common. Particularly the local public transport modes *Bus* and *Tram* get mistaken surprisingly often for Cars or Trains.

Table 4.6 reveals that while the most common modes are hardly affected at all by the spatial distortion, the already quite poorly classified modes suffer particularly strongly.

	Car	Bike	Walk	Train	Tram	Bus	Precision
Car	95%	48%	8%	6%	5%	52%	90%
Bike	0%	31%	1%	0%	3%	0%	86%
Walk	2%	17%	90%	1%	14%	8%	87%
Train	3%	3%	0%	94%	2%	1%	95%
Tram	0%	0%	0%	0%	76%	1%	78%
Bus	0%	2%	0%	0%	0%	38%	75%
Recall	95%	31%	90%	94%	76%	38%	

Table 4.7: Confusion matrix with no spatial distortion for the pure classification problem.

	Car	Bike	Walk	Train	Tram	Bus	Precision
Car	94%	50%	8%	7%	8%	53%	86%
Bike	0%	25%	0%	0%	4%	0%	88%
Walk	3%	22%	91%	1%	25%	18%	83%
Train	3%	2%	0%	92%	2%	2%	95%
Tram	0%	0%	0%	0%	61%	1%	75%
Bus	0%	1%	0%	0%	0%	26%	71%
Recall	94%	25%	91%	92%	61%	26%	

Table 4.8: Confusion matrix with large spatial distortion for the pure classification problem.

Classification based on ground truth segmentation is unsurprisingly much better than if it is based on inferred segments, as seen in Table 4.7. In particular the walking segments benefit greatly. While Trams now are decently discovered, buses and bikes still suffer from poor recall values, even if the precision has improved significantly.

In terms of spatial distortion, the deterioration is less dramatic than for classification on inferred segments, as can be seen in Table 4.8. While it is still the modes with poor recall that suffer the most, the decline is smaller than before.

4.4 Discussion

4.4.1 Overall results

On the most important question, concerning what quality the data from passive tracking would need to deliver in order to allow traffic mode detection, the following can be observed:

Limited benefit from high temporal granularity

In terms of temporal granularity, sampling at too high a frequency will not benefit the classification results, and on the contrary even deteriorate them in the schemes tested here. In addition there is no significant interaction with spatial accuracy and the above holds for all tested values, as seen in Figures 4.6 and 4.7. On the one hand, this confirms findings from the literature (Bolbol and Cheng 2010) that claim an ideal sampling rate in the order of magnitude of about a minute, but on the other hand it generalises them to the case where spacial accuracy of the measurements cannot be treated as a given.

Spatial accuracy is always helpful

In terms of spatial distortions, the image is less clear. The traffic modes that made up the bulk of the data and that were well classified in the absence of spatial distortions continued to be correctly identified most of the time even for the largest spatial distortion that was tested. However, the other modes, that already were poorly identified in the base case, suffered considerably under spatial distortions. It seems likely that some of the poor results for local public transport can be attributed to the labelling scheme that did not allow users to insert missing trip legs. The fact that many bus/tram trips comprise the access/egress walking trip legs as well means that the classifiers cannot reliably learn that bus/tram legs start at corresponding stations, which would explain the low added value of the GIS features. In addition, as slow segments carry a local public transport label, the classifiers can no longer reliably learn that slow speeds are indicative of walking segments, which would explains the bad performance on walking segments compared to the literature. The breadth of users targeted in the original data collection campaign thus came at a rather significant cost in classification quality.

On all spatial distortions that were tested, there was no complete breakdown of the methods on the bulk of the data, i.e. on the *car* and *train* modes. Rather there was a steady decline from the baseline. This means that there is no clear minimum uncertainty (in the range that was tested) beyond which detection becomes completely unfeasible. But clearly, the more accurate the data, the better the results. This is in contrast to the temporal granularity, where too much detail could be detrimental.

Segmentation + RF yields best results

In terms of the methods compared, it became clear that overall, the best approach seems to be to apply a decent segmentation and classify based on segments. Smoothing is not necessary when segmenting first, but absolutely necessary if the classification happens in a pointwise fashion, as seen in Figure 4.4. Random Forests had stable and qualitatively appealing results.

4.4.2 Sensitivities

The results of the sensitivity analysis conformed to expectations, at least qualitatively.

GIS effectiveness less than expected

GIS information does contribute to the classification, as shown in Table 4.4, although not quite to the degree expected from the literature (Stenneth et al. 2011). If one looks at the feature importance for tram and bus, GIS features rank among the most important predictors. Interestingly this is not the case for train, where the instantaneous (calculated) speed at the end of a segment as well as overall displacement are the top features. With very few GPS fixes inside trains, the speed at the beginning/end comprise more or less the trip as a whole, as does displacement, since any stops in the middle cannot cause a segment break, leading to larger segments than are observed for other modes. Thus, for trains, GIS information may not be adding much in cases with data such as that used here, where there are few or no valid GPS fixes. Thus, as data was skewed away from the transport modes where GIS was helpful, the overall contribution was limited.

With respect to segmentation thresholds, having thresholds that result in more segments, as the one shown in the results, can lead to oversegmentation and hence to high edit distances. While smoothing can remedy some of this deterioration (as shown in the Supplementary Material only), it does not lead to results that beat that of the baseline.

Using a cross-validation scheme that cross-validates by trip instead of by user, the results get slightly better, but mostly the standard deviation decreases substantially. The folds used in this scheme do not fundamentally differ from one another, since all folds contain trips from all users, leading to very stable, but overly optimistic results, underestimating the uncertainty in the quality measures when generalising to people that did not contribute to the training data.

Importance of good segmentation

The results from the ground-truth segmentation sensitivity analysis underline the importance of good segmentation. The closer the segments get to stages, the better the expected results. It also shows that the good results reported on pre-segmented trips should not be used to form expectations about the classification accuracies in situations when the stages are not given.

4.4.3 Confusion matrices

Overall, even with features limited to those available to passive tracking schemes, the overall accuracies were in the range expected from the literature. However, some modes were quite poorly identified.

Reasons for bad performance on certain modes

For the Walk label, this can partly be explained by labelling issues discussed in Section 4.2.1: There were plenty of very slow segments in bus stages during training and therefore, while all seen walking stages are slow, not all slow segments that should belonged to walking stages.

Regardless of the segmentation used, bicycles were not that easy to detect. They seem to take some place between walking and cars. This appears plausible, as a bicycle leg can look almost as one on foot if it is uphill, or can be nearly indistinguishable from a car in city traffic, if it shares the same restrictions in terms of traffic lights, stop signs and the like. While this distinction is easier when accelerometers are available, distinguishing the three modes is much harder based on GPS alone.

Results qualitatively similar on GT segments

The classifications based on the ground truth stages can help to shed at least some light on the effect of the less than perfect labels. To be clear, the effect is confounded with the fact that the problem of labelling stages rather than segments is easier, but it is possible that there are some pointers nonetheless. Most striking is the increase in quality for the walk stages. The deterioration of the results that derives from the deterioration in spatial accuracy is still visible, but less extreme than for the results on the move segments.

As slow segments are no longer seen in isolation (as walking stages were often merged into stages of other modes), significantly more of the slow segments actually reflect walking stages, which makes the stages labelled *Walk* significantly more separable in the feature space.

Again, as for the inferred case, Trams are more easily identifiable than buses, due to the fact that there are fewer cities in which there are trams in the first place, making the GIS information more useful here (as reflected in the higher feature importance). The buses are still hard to separate from cars, but this does not come entirely surprising, since they do share similar characteristics.

For the modes that were detected well, the deterioration was similar using the stages as segments as it was for the results to the complete problem of combined segmentation and classification.

4.5 Conclusion and outlook

In this case study, commonly used methods for the classification of traffic modes to information that could be available from passively sensing mobile phone data through the mobile phone network have been tested for various levels of temporal granularity and spatial uncertainty. Realistic data from over 130 users collected over half a year, which was annotated by those users was used in the study.

Car and Train stages are discernible

For the feasibility of purely passively sensed transportation surveys the results reported here seem to indicate that with the data used in this study, they are feasible for separating *Car* and *Train* trips with accuracies in the orders of magnitude found in the literature (Shen and Stopher 2014), even for the combined segmentation and classification problem. The thresholds that are required for achieving this are temporal granularities in the order of 30 to 60 seconds, as previously stated in the literature for spatially accurate data, (Bolbol and Cheng 2010) and a spatial uncertainty of less than 200 m in each direction, which is within reach of today's passive mobile phone positioning technology.

Temporally, order of minutes is sufficient

Temporally , ever increasing granularities are not necessarily beneficial. The stronger the spatial uncertainty, the more high sampling rates deteriorate the results. Sampling at a rate in the order of magnitude of one minute seems to be ideal, with slightly sparser samples not deteriorating the results too much.

Spacially, less error is always better

In terms of spatial uncertainty, more accuracy always leads to better results, irrespective of the temporal granularity. The effects of spatial uncertainty are strongest for pointwise classifiers that do not include some form of smoothing of the results.

As for the most successful strategy, the best results were obtained by segmenting a trip into meaningful parts and classifying based on segmentwise features. Random forests have yielded the best overall results in this setting.

Diversity of the people in the data

Having a very wide range of people contributing their data, distributed across a wide range of geographical situations over such a long time came at the cost of a reduced interaction when labelling the data, leading to some fused stages. While this leads to a more representative dataset than what can be collected with just a few dedicated researchers (Feng and Timmermans 2016; Nitsche et al. 2014; Stenneth et al. 2011), the quality of the data is harder to control. It is not possible to tell how strongly this problem affected the data quality, but the results shown here can be seen as a lower bound of what is possible, and confidently separating more modes may be possible with more accurate labels.

Need for good segmentation and benchmark datasets

However, traffic mode detection based on passively sensed data is not yet satisfactorily solved. In particular there are two areas where there is need for additional work. The first is finding segmentations whose resulting segments are closer to stages. As indicated by the leap in classification quality observed when using ground truth segments, there still seems to be potential. In addition, the different benefits from having a very accurately labelled, strongly controlled, maybe even balanced data set for training typically used in research versus the breadth of different settings and locations afforded by the data set used here could be investigated to inform the collection of a training data set to minimise generalisation error.

While, as just stated, the problem of mode detection from passively sensed data cannot yet be viewed as solved, even with those first steps into that direction, thinking about what will be possible with this kind of information becomes a possibility. One such option is the short term prediction of Eulerised movement data which is covered in the next section.

Chapter 5

Predicting single mode traffic flows

Was soll mir das Lob von Menschen, welche nicht tadeln können?
— Annette von Droste Hülshoff

This chapter is based on a research article that was submitted to *IET Intelligent Transport Systems* under the title 'Architecture Independent Residual Deep Learning for Traffic Flow Prediction' and is currently under review.
The scientific contributions of other researchers to this chapter were the following: Robert Weibel provided supervision and helped prepare the manuscript for submission.

5.1 Study setup

Consequences of large scale mode detection

As was seen in Chapter 4, modes of transport can become identifiable even from data that is passively gathered by mobile phone companies. It is conceivable that they could thus create a system providing a very detailed, multimodal picture of traffic, similar to the one Google already provides for the streets, but including precise numbers, mode transitions at every station and so on. One natural question to be asking such a system is about the traffic conditions of the near future.

This shift would amount do a decoupling of the sensor system from the infrastructure. As the sensors in this brave new world would be with the people, questions can be asked even in regions where an infrastructure based sensor grid is not available and entirely new questions are possible. This flexibility in the questions asked means that the stability of sensors mounted on slow changing infrastructure cannot be relied upon any longer. As movement flows in general can be more volatile than those of e.g. arterials, a certain flexibility of the methods may be required.

Thrust perpendicular to architecture

In the literature on short term traffic prediction, as in many other fields in recent years, deep learning has become a popular tool. Deep learning is known to require massive amounts of data (Sun et al. 2017; Vlahogianni, Golias and Karlaftis 2004); Amounts that cannot simply be assumed even for infrastructure mounted sensors that record relatively stable traffic flows. The focus of research so far has been on the architecture of the models (cf. Section 2.3) and some fairly large models with large numbers of parameters have been fitted. In all of this the incorporation of domain knowledge has not been front and centre of those efforts. This situation is diametrically opposed to the situation described in Chapter 3, where very simple (rule based) models have been fitted using an abundance of domain knowledge.

Contributions

In this chapter, the focus therefore lies on how domain knowledge can be used in traffic flow predictions for different architectures of deep neural networks. In detail the contributions are:

- A restatement of the traffic flow prediction problem into a two stage process allowing for the incorporation of domain knowledge.
- An analysis of how existing ways to incorporate domain knowledge into deep learning architectures impact prediction results
- A demonstration of how implementing the proposed two stage process can reduce the prediction error even further than systems found in the literature.

The problem that this chapter focuses on is the regression problem discussed in Section 2.3 where future values of traffic flow are being predicted using observed (and therefore past) values of the flows. The chapter starts with a formalisation of the problem to be solved, to make it as explicit as possible and to have a reference against which the proposed alternative can be compared, which happens in the subsequent section. Finally, a case study demonstrates the merits of the proposed improvements.

5.2 Traditional problem statement

This section presents the problem that has to be solved. The value of carefully stating the problem first lies in contrasting it with the restatement later. While the problem remains the same, the proposed way of solving it follows very naturally from the restated version.

5.2.1 Traffic flow as a time series

Let $\mathfrak{S}$ denote the set of all available stations and $\mathfrak{T} := 1, \ldots, T$ a set of equidistant points in time. Then the value of traffic flow at the station $s \in \mathfrak{S}$ at time $t \in \mathfrak{T}$ is denoted by $x_t^{(s)}$. For the most frequently addressed univariate case with only one station at a time, the problem is the following: Given $(x_u^{(s)})_{u=1,\ldots,t}$, predict the values $x_{t'}^{(s)}$ for some $t' > t$ usually with a fixed difference between t and t'.

The first extension to this problem is that to the multivariate version that given $\left(x_u^{(r)}\right)_{u=1,\ldots,t;r=1,\ldots,N_S}$ for some $\mathfrak{S}' \subset \mathfrak{S}$ and $N_S := |\mathfrak{S}'|$ one wants to predict $\left(x_{t'}^{(r)}\right)_{r\in\mathfrak{S}'}$. Here, $\mathfrak{S}'$ is considered to be an ordered list and the stations are denoted by their position in that list.

Some authors trained multiple models to solve the above problem for different lookahead periods $t' - t$ (Lv et al. 2014), but it is of course possible to pose the problem as to predict several time steps at once. Therefore a fixed time horizon $\Delta t \in \mathbb{N}$ was chosen and the information at all time steps $t+1, \ldots, t+\Delta t$ was predicted given the information at the last known state t. In addition, a set of covariates $\mathfrak{C}$ was used of which at time t also the values at times $t+1, \ldots, \Delta t$ were known or at least reasonably easily and accurately predictable, such as the weather. Their values are denoted by $c_t^{(i)}$ for $i \in \{1, \ldots, N_C\}; t \in 1, \ldots, T$ where $N_C := |\mathfrak{C}|$. Most methods for traffic flow prediction do not strictly speaking use all available information at time t for the prediction but instead choose a fixed window of size w and only use the information of the time steps $t-w+1, \ldots, t$. This convention will be followed and thus the problem can be stated as follows: Given

$$\begin{bmatrix} c_{t-w+1}^{(1)} & \cdots & c_{t-w+1}^{(N_C)} & x_{t-w+1}^{(1)} & \cdots & x_{t-w+1}^{(N_S)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c_t^{(1)} & \ldots & c_t^{(N_C)} & x_t^{(1)} & \ldots & x_t^{(N_S)} \end{bmatrix} \tag{5.1}$$

and

$$\begin{bmatrix} c_{t+1}^{(1)} & \dots & c_{t+1}^{(N_C)} \\ \vdots & \ddots & \vdots \\ c_{t+\Delta t}^{(1)} & \dots & c_{t+\Delta t}^{(N_C)} \end{bmatrix} \tag{5.2}$$

predict

$$\begin{bmatrix} x_{t+1}^{(1)} & \dots & x_{t+1}^{(N_S)} \\ \vdots & \ddots & \vdots \\ x_{t+\Delta t}^{(1)} & \dots & x_{t+\Delta t}^{(N_S)} \end{bmatrix} \tag{5.3}$$

5.2.2 Evaluation criterion

The quality of the multivariate prediction can be assessed in various ways. A variety of different measures has been used, such as mean absolute percentage error (MAPE) (Xie and Zhang 2006), mean absolute error (MAE) (Vlahogianni, Karlaftis and Golias 2005), and root mean squared error (RMSE) (Sun, Zhang and Yu 2006). Traffic flows can be as low as zero and any measure that divides by the true data (such as MAPE) is therefore inadequate for two reasons. The first is the immediate problem of division by zero, that can happen if there are no vehicles on the road segment during an observed period. The second problem is that the errors will tend to be largest for small ground truth values, so the model will focus on correctly predicting small values. However, most applications are probably mostly interested in the accurate prediction of hours where there is a large flow and hence a measure that does not disadvantage those flows is preferable.

MAE and RMSE both avoid this problem by only looking at the magnitude of the deviation, irrespective of the target value. Applications can be assumed to be interested in avoiding gross errors which are penalised harder by RMSE. Therefore this measure was used.

While optimising, all possible time horizons were used for predictions equally and the following expression was minimised:

$$\text{RMSE} := \sqrt{\frac{\sum_{i=1}^{\Delta t} \sum_{j=1}^{N_S} \left(x_{t+i}^{j} - \hat{x}_{t+i}^{j}\right)^2}{N_S + \Delta t}}.$$

However, to show the evolution in time of the prediction error also the error rates for every prediction horizon was calculated:

$$\text{RMSE}_{t'} := \sqrt{\frac{\sum_{j=1}^{N_S} \left(x_{t+t'}^{j} - \hat{x}_{t+t'}^{j}\right)^2}{N_S}}, \quad t' \in \{1, \dots, \Delta t\}.$$

5.3 Two stage estimation and residual problem statement

As already alluded to, the motivation behind this work is the fact that in transportation science the number of training cases are limited. This in turn limits the size of the model that can be used and thus the quality of the results. The idea is to make the prediction problem easier for the neural networks by incorporating domain knowledge.

In the traditional problem that deep learning models on traffic prediction have to solve, they have to discover the diurnal and hebdomadal patterns on their own. However, knowing what they reflect, it is possible – and intuitively it would seem helpful – to relive the models from that burden, by handing those patterns to them.

5.3.1 Two stage estimation

Some researchers have provided information from a day and/or a week ago (Wu and Tan 2016) as covariates in their models. However, giving only the previous week's value can be considered rather simplistic. After all, many weeks may be special due to bank holidays or popular events such as fairs or concerts. In weeks after those special occasions, using the special week's data as covariates may be of limited use.

Therefore it makes sense to calculate still simple but more robust estimation as a first stage and use it as a feature in the second stage, the actual prediction. Simplicity means that the covariates can be provided without actually learning an entire model or similar, such that they really are just covariates, while robustness ensures that events of limited duration do not disturb that signal.

One very robust way of achieving this would be to use global daily and weekly averages, which would certainly get rid of effects from one-time events in the preceding week of a data point, but on the other hand would also fail to capture other seasonal effects, such as a difference between summer and winter which may be present.

Robust and simple first stage

To capture both, the proposition here is to use a robust measure of centrality over a moving window of several weeks, specifically the median over the last 3 weeks. The robustness of the centrality measure ensures the robustness of the introduced covariates. The moving window can be enlarged to reduce variance at the cost of increased bias.

The covariates calculated in this way can simply be added to other covariates that may be used in the prediction. Alternatively they can be used to replace the covariates based on the preceding week.

Either way, the problem from above requires an additional step to be calculated before solving the problem from Section 5.2.1, resulting in a two-stage process for short-term traffic flow prediction.

5.3.2 Residual problem statement

In addition to providing more robust covariates, there is a second avenue to improve prediction accuracy: simplifying the problem. In time series analysis the removal of the systematic components is a well established practice to make a process stationary (Shumway and Stoffer 2017). In deep learning, while in a different context, it has also been found that it can be beneficial to only model deviations instead of the signal itself. This approach, known as Resnet (He et al. 2016) has been very successful, in particular in deep architectures. In simple end-to-end approaches on deep learning found in the literature, this is not widely used.

The following adaptation of those two ideas to short-term traffic prediction will be used here: Let there be a base model, which for time t and station s predicts the value $\hat{x}_t^s$. Further, let the residual of that prediction be defined as $r_t^s \coloneqq x_t^s - \hat{x}_t^s$. Then rephrase the problem as finding

$$\begin{bmatrix} r_{t+1}^{(1)} & \dots & r_{t+1}^{(N_S)} \\ \vdots & \ddots & \vdots \\ r_{t+\Delta t}^{(1)} & \dots & r_{t+\Delta t}^{(N_S)} \end{bmatrix} \tag{5.4}$$

given

$$\begin{bmatrix} c_{t-w+1}^{(1)} & \dots & c_{t-w+1}^{(N_C)} & r_{t-w+1}^{(1)} & \dots & r_{t-w+1}^{(N_S)} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ c_t^{(1)} & \dots & c_t^{(N_C)} & r_t^{(1)} & \dots & r_t^{(N_S)} \end{bmatrix} \tag{5.5}$$

and

$$\begin{bmatrix} c_{t+1}^{(1)} & \dots & c_{t+1}^{(N_C)} \\ \vdots & \ddots & \vdots \\ c_{t+\Delta t}^{(1)} & \dots & c_{t+\Delta t}^{(N_C)} \end{bmatrix} \tag{5.6}$$

Note that even when framing the problem this way, the actual values of the flow can be taken as a set of N_S covariates, reversing the traditional roles of the signal and the covariate. While this increases the number of parameters to fit, it may prove beneficial in cases when deviations from the base models are predictable based on the progress along the diurnal pattern.

As the base model, the robust estimator motivated in Section 5.3.1.

5.4 Case Study

In this section, the respective merits of the two stage solution and the rephrasing of the prediction task using the residual problem statement are to be evaluated. A lot of care will be given to making the comparison fair by establishing a strong baseline against which to test.

5.4.1 Baseline models

ARIMA

There are several layers of benchmarks that are given as comparison. First the error rates of pure historical averages and ARIMA based predictions are given to provide a comparison with formerly very popular models. While they have extensively been shown to be inferior to deep learning based approaches (Lv et al. 2014; Ma et al. 2015; Shao and Soong 2016; Tian and Pan 2015; Zhao et al. 2017), they are shown here to put the improvement from solving the residual problem into perspective. Also, the historical values given are those based on the robust estimation from the past in order to allow for an isolation of the second stage of the two stage model.

FFNN and LSTM

In addition to the traditional benchmarks, deep learning based benchmarks will be established. The two most commonly used networks are the feedforward and the recurrent type, as discussed in 2.3.4 and hence both will be used as benchmarks. To ensure that any improvement is actually due to the two stage prediction and/or the restatement of the prediction problem, great care is given to provide good benchmarks. As many factors can influence the performance and an exhaustive search over all parameters would be prohibitive, the optimal models for the benchmark models are established by optimising groups of hyperparameters sequentially.

Capacity

First, a hyperparameter search on depth and number of hidden nodes is performed, as is fairly common in the literature. This should provide an impression on the general size of models that can be used in this context. In the literature, sometimes very small models with hidden layers with tens of units are used (Shao and Soong 2016), mostly with univariate predictions. Here only somewhat larger models with sizes of layers in the hundreds are investigated, especially since the prediction is multivariate, necessitating somewhat larger capacities.

For the depth of the model, one to three layers are tested in the grid search and as for the width, 256, 512 and 768 hidden units per layer are taken. This is in the order of magnitude that can also be found as the successful choice in the literature (Zhao et al. 2017).

Dropout

In a second step, and this is not commonly discussed in the literature, a search for the optimal dropout rate is conducted. Dropout is a technique for regularisation. As the number of parameters to be fitted is fairly large given the number of hidden units and layers used and the training data are limited, there is the danger of overtraining. This means that the capacity of the model is large enough to allow it to not only learn the real connections between the different variables but also the idiosyncrasies of the training data set. This typically leads to bad generalisation errors. Dropout (Srivastava et al. 2014) is one technique to reduce it. It works by setting certain neurons to exactly zero during training. This forces the network to learn patterns redundantly, since it cannot rely on any neuron to perform its duty. These redundant patterns are typically more robust. The dropout rate determines the probability with which a neuron is set to zero during training. The higher the rate, the more robust the estimator needs to be, whereas a low rate yields results close to what can be obtained without dropout. For the grid search the values of 40%, 30%, 20%, 10% and 5% are tested.

Covariates

The third step of establishing the benchmarks consists of using different sets of potentially useful covariates. The first set of covariates consists of time related variables. The reasoning behind it is that there are clear weekly patterns that are known to exist. The prediction may vary depending on whether a given past development of traffic flow happened on a Sunday or a Monday and whether it was observed on a morning or towards the evening.

Therefore the time of day is added as a linear variable between 0 and 1, taking small values shortly after midnight and large values close to the following midnight. In addition, seven one-hot dummy variables for the day of the week are added to indicate the weekly cycle. Lastly in this first batch of covariates a variable that takes the value of 1 on midnight and zero otherwise was used. This may be helpful for the forget gates of the LSTM cells, as any information that has to be stored daily can easily be flushed this way.

The second group of covariates is added to illustrate that also thematic covariates can be used seamlessly in this setup and to incorporate the findings of Elhenawy and Rakha (2016). Weather data are chosen, as there is an intuitive connection between the weather and traffic and it is freely available for the study area through the NOAA website (Department of Commerce 2017). There are several ways in which the additional information can be added. One option is to simply add every sensor and every value that sensor provides separately. However, the data come at a temporal resolution of one day, so are correlated. Therefore summary statistics (average plus quartiles one and three) were used for each of the following variables: average wind speed (numeric), precipitation (numeric), average temperature (numeric), maximal and minimal temperature (numeric), peak 2-minute and 5-minute wind gushes (numeric), fog (indicator), heavy fog (indicator) and smoke or haze (indicator). The same values for all measurements of the day are used. Here, the dummy variable marking the beginning of a day may be helpful for the LSTM, as it can signal to forget the weather data of the past day, if this helps prediction.

The predictability of the weather is obviously not a given, as it is not known in advance what e.g. the total precipitation of a day will be. However, one day forecasts are of high quality today so that in an operational system, those forecasts could be used instead of the real numbers.

Measuring uncertainty

All runs in this first stage are run five times and the choice is based on the median values obtained from them. This limits the effect of the randomness inherent in deep learning and allows for a more robust assessment of the effects of the different parameters.

5.4.2 Measuring the effect

To reason about the effect of the two stage prediction procedure and the residual problem statement, all possible combinations of the two are tested. The reason why this is not done in stages comes from possible interactions between the two. Any combination might be particularly useful for analysis. The following possible settings were used in all sensible combinations:

- *Last week*: Whether or not last week's values were shown during training. Part of this was to benchmark the more robust estimation of the simple predictor against what was traditionally used in the literature. In addition, there may be interactions with the residual problem statement.
- *Show raw values*: Whether or not the traffic flow number should be visible in training. These are the ones where the strong daily and weekly pattern are visible and the values that ultimately are of interest.
- *Show residuals*: Whether or not the residual values should be visible during training. At least one of the raw or the residual values needed to be present, as otherwise the fit would be an estimate on time of day and day of the week alone, which cannot be expected to be useful.
- *Fit raw values*: Whether the raw values should be fitted (standard approach) or whether the residuals themselves should be fitted.

Note that most hyperparameters from the benchmark models are taken and there is no second phase grid search for all the combinations of the variables just described. The optimal values from the traditional problem statement are retained. This way any observable improvement is real and even a conservative estimation of the difference. There is one exception to this pertaining to the capacity of FFNN models. With the added bands of information, the input space grows significantly, so a model with larger capacity was fitted to ensure the model can handle the additional information.

To also ensure that the results are not merely the result of a lucky constellation of pseudo-random numbers drawn during the learning process, all runs in this phase were also repeated five times. This allows a better assessment of the changes in the results due to the model itself.

5.4.3 Data and implementation

Data source

The PEACROMS dataset on the Caltrans district 07 (Los Angeles) was as the basis for the experiment (PeMS 2017). This district has a large proportion of stations that have been active for a long time and that have just very few missing values. The few missing values that do appear are filled by day-of-the-week and time-of-day averages that are affinely transformed to match the start and end points of the gaps. The five minute information was aggregated to fifteen minute intervals as often observed in the transportation literature, although for both methods it does not really matter. Taking five minute numbers directly instead of the 15 minute aggregates would increase the noise and the memory cells would have to remember information for more time steps, but LSTMs have been shown to work for over a thousand time steps (Hochreiter and Schmidhuber 1997), which would make them seem adequate even in the five minute setting.

The differences in traffic flow between stations are dominated by the number of lanes, as every station reports the sum of all traffic flow over all lanes at its location. Therefore, the raw traffic flows are divided by the number of lanes to get the data into similar ranges.

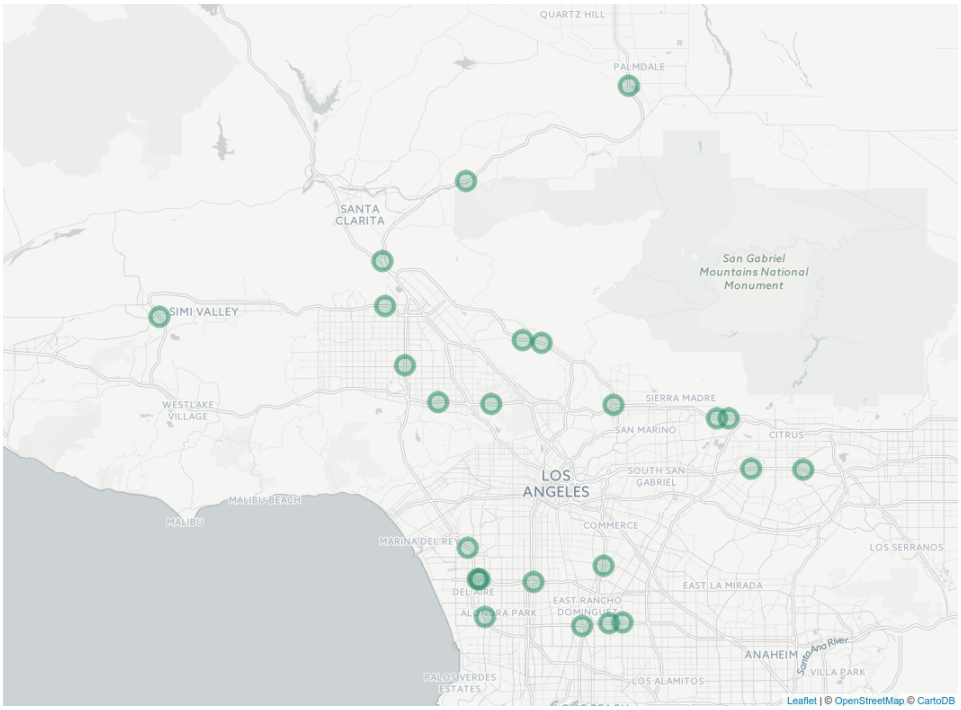


Figure 5.1:
Map of the chosen stations.

The complete years 2014-2016 were downloaded and prepared. The first two years are used for training, the first half of 2016 as validation data set and the second half of 2016 as test set. Convergence is decided on the basis of the validation data only while all reported values are from the test data. The final retained state of the parameters is taken from the point in training where the validation error was lowest. As this would be too optimistic of an estimate for generalisation, the retained model is then used on the test data and only the results there (i.e. true generalisation) are reported.

Stations used

A subset of thirty stations was used. The stations were chosen by regularly sampling from the list of stations that was ordered by quality of the simplest benchmark solution for those stations in order to avoid both an overly favourable as well as an overly unfavourable presentation of the proposed solution. The map displaying the chosen stations can be seen in Figure 5.1.

Implementation

The models were implemented on TensorFlow[1]. The code can be found on a GitHub repository[2] and is publicly accessible.

All the calculations were done on a GPU cluster where every node is equipped with an Nvidia Tesla K80 graphics card. Every combination of parameter settings was run on only one node, so the advantage of having a cluster was being able to run multiple parameter settings simultaneously. This was especially helpful as all the runs had to repeated multiple times to allow for an estimation of the randomness inherent in the final results.

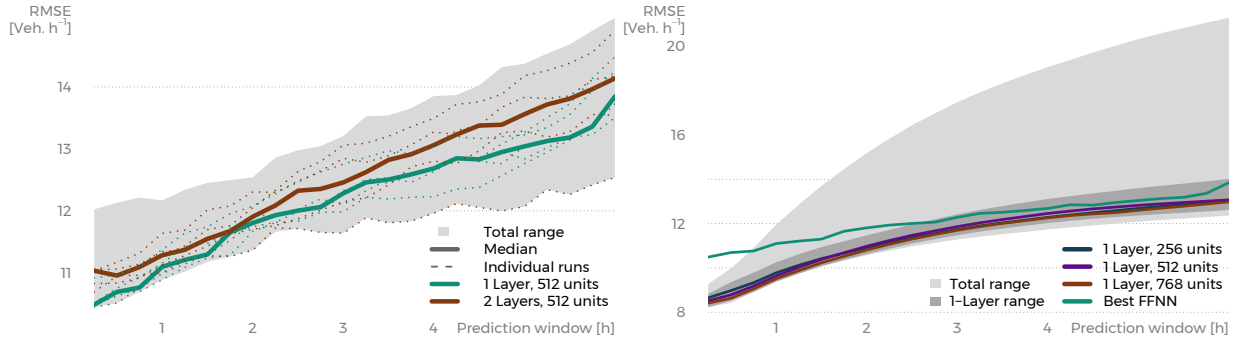
[1] www.tensorflow.org
[2] https://github.com/o1i/traffic_prediction.git

5.5 Results

The results presented are first step by step the established benchmarks, followed by those of the two stage process and the residual problem statement.

The benchmarks of using past values and ARIMA models are not shown graphically here, as they were much worse than the deep learning based models. Unsurprisingly, using past values leads to a constant error rate for all prediction horizons, since the past values do not change whether the forecast is for fifteen minutes or two hours. The RMSE was at 44.15 and therefore well above any of the other solutions tested. The benchmark of using an ARIMA on those predictions led to lower values that included the behaviour of better predictions in the short term than for the ones further out, but still the RMSE started at 22 and rose to almost 38 vehicles per hour and lane, which still is in the order of magnitude of twice the error rate of the other approaches. Clearly, using more sophisticated models is an improvement over the simple predictors.

Figure 5.2 presents the main results of the grid search of establishing the benchmarks. As the grid search produced significant amounts of results, only the most relevant are retained in the figure. The different ordinates are chosen deliberately to illustrate the important effects discussed in Section 5.6.



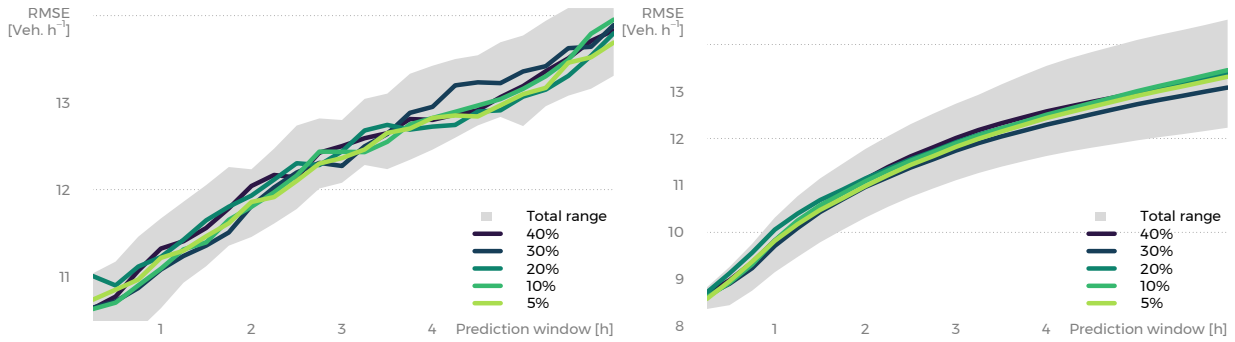
(a) FFNN (b) LSTM

Figure 5.2: Results for the grid search on number of hidden units per layer and number of layers. To the left, all tested parameter settings had a large overlap over the repetitions. The two parameter settings that are shown reflect the best one overall and the one that had one run with very good results for longer prediction horizons. To the right the results for the LSTM networks are shown. To illustrate the reduced overlap, the range of the results for the one layer models are emphasised and for a comparison with the FFNN models, the best parameter setting overall from the left hand side (1 Layer, 512 units) was included.

The results from the dropout stage of the benchmark production can be found in Figure 5.3. Again, FFNN results are to the left and LSTM results are to the right.

The results from the analysis of the covariates can be found in Figure 5.4. Again, FFNN results are to the left and LSTM results are to the right.

Finally, the results from the various combinations of the two stage process and the residual fitting can be found in Figure 5.5.



(a) FFNN (b) LSTM

Figure 5.3: Results for the dropout stage of the benchmarking. The gray polygon denotes the range of results over all runs whereas the coloured lines represent the medians over a fixed dropout rate. On the left hand side are the results of the FFNN models whereas the results from the LSTM nets can be found on the right. The scales of the ordinates differ to emphasise the difference *within* the different architectures rather than those *between* them.

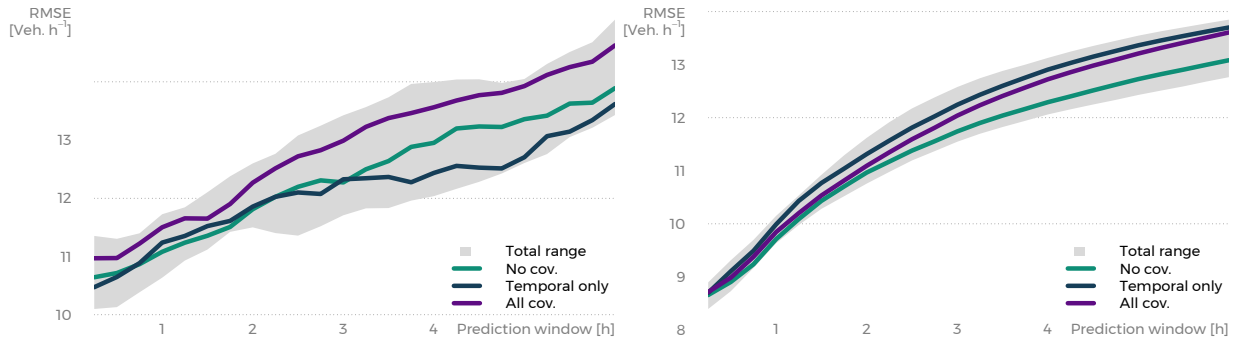
5.6 Discussion

Grid search

The first stage of the grid search produced qualitatively different results for the two tested architectures. For the FFNN networks seen in Figure 5.2a there was significant overlap between the different parameter settings. The uncertainty from the different runs (as measured by the range over 5 realisations) was larger than the differences between the medians for the settings. Nonetheless, there were better and worse settings and overall the best was the one with simply one layer and 512 hidden units in that layer. This may come somewhat as a surprise given the fact that the curve of that parameter setting ends up somewhere in the middle of the range at the longest time horizons. This is mostly due to one outlier produced by the parameter setting using 2 layers and 512 hidden units, that in the other four cases performed worse than the overall best parameter setting. Other parameter settings that performed well for the longest prediction horizons were from settings with even more layers which performed even worse in the short and medium prediction horizons. Therefore, for the next steps, the parameter setting with 1 layer and 512 hidden units was chosen for FFNN networks.

For the LSTM networks on the other hand presented in Figure 5.2b, the picture was different. The more layers that were present, the worse the results got, especially for the longer prediction horizons. This is illustrated by highlighting the very narrow band of the one layered models and the very close proximity of all the medians of the one layered models. Only one other parameter setting – the one using 2 layers and 768 hidden units per layer – had a large overlap with the one layered models, but was inferior when averaging over time. Therefore of those four parameter settings that produced very similar results, the most parsimonious was chosen and for all further steps, the LSTM networks all were given 1 layer and 256 hidden units in that layer.

A comparison between the two architectures on the grid search reveals that FFNN networks seem to be inferior to LSTM networks, particularly in the predictions that are close to the last observed data point, which can be considered most important.



(a) FFNN (b) LSTM

Figure 5.4: Results for the covariate stage of the benchmarking. The gray polygon denotes the range of results over all runs whereas the coloured lines represent the medians over a fixed set of covariates. Either no covariates were used, only the block of the temporal covariates or all covariates, including both temporal and thematic ones. On the left hand side are the results of the FFNN nets whereas the results from the LSTM nets can be found on the right. The scales of the ordinates differ to emphasise the difference *within* the different architectures rather than those *between* them.

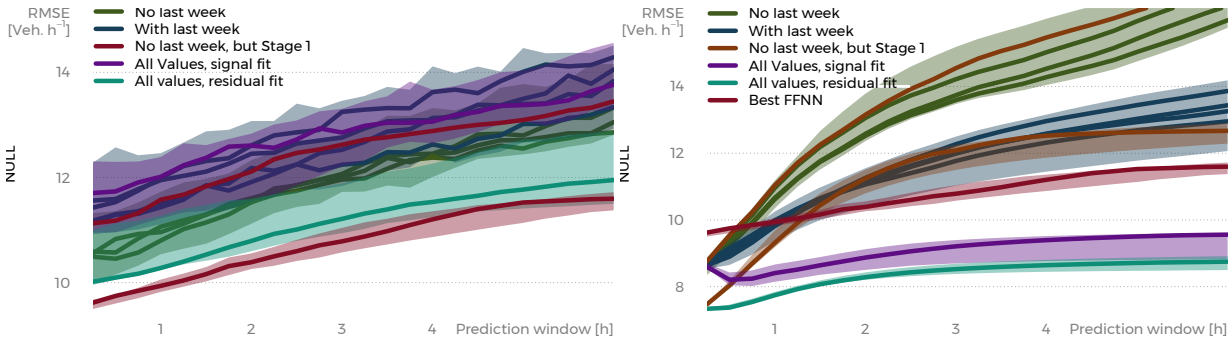
Dropout

In terms of dropout, both types of models showed near indifference in Figure 5.3 to the precise level and therefore almost any of the tested values could have been retained. For both types of networks, 30 % was retained. In the case of FFNN because of its relatively good performance in the first two hours, whereas for the LSTM models, as it seems to be slightly superior over many of the prediction horizons used.

Covariates

In terms of covariates, again the FFNN networks showed large overlap and no entirely clear picture, as demonstrated by Figure 5.4a. In the short run, the covariates do not seem to be able to contribute to prediction, and the variant without covariates is even slightly superior to using the temporal block of covariates. However, slightly further out in the prediction, the temporal covariates could help contribute to the prediction. This is intuitive, as in the very short term, the current situation almost fully determines the future traffic, but very far out the current traffic situation is meaningless, thus other factors can start being helpful somewhere in between. Therefore, in the following, the temporal covariates were used in addition to the pure flow values.

For LSTM networks, the situation was different as illustrated by Figure 5.4b. The runs are relatively well separated between those that use covariates and those that do not, with the latter showing lower errors. This runs somewhat counter to other attempts at incorporating weather conditions Elhenawy and Rakha (2016) but may be a result of different data that was used.



(a) FFNN (b) LSTM

Figure 5.5: Results for the two stage procedure and the residual problem statement. Deviating from preceding figures, ranges of groups of parameter settings are now coloured differently to allow the differentiation between actual effect and random noise. One exception are the runs that use the raw values and the ones from stage one but not last weeks. For those two there is a significant difference whether the raw values or the residuals are predicted and thus they are plotted separately. The FFNN models on the left hand side were fitted with a larger capacity than the benchmarks, which was necessary due to the increase in the number of features.

Two stage model

Both models could benefit from the two stage model, but differently. If fitted with the parameter settings from the benchmarks (not shown here), the FFNN models seem to suffer from capacity limitations: With one exception, having more than one band of information decreases performance. This includes last week's value which is very commonly found among the features of traffic flow prediction. As many researchers have used this feature successfully in a variety of different settings (Chandra and Al-Deek 2009; Kamarianakis and Prastacos 2003; Wu and Tan 2016), this points to a potential capacity problem with the network. More hidden units may be require to handle the additional information. This can partially be remedied by increasing the capacity of the model, shown in Figure 5.5a. With higher capacities, the overall error is markedly lower. Still, however, additional bands of information typically lead to a deterioration in performance. Using the first stage as an additional band was no exception to this rule in most cases. However, in the most important case – the combination with the residual problem statement – the first stage could help reduce the input complexity and lead to better prediction accuracies for very short term predictions.

For the LSTM model shown in Figure 5.5b, the effect was very clear. The two combinations that stand out – the purple and the turquoise line – both use all three sets of parameters: raw values, last week's values and the results from stage one. Any combination that did not use all three fared significantly worse for prediction horizons of more than an hour. While in the very short term there was not a dramatic change, for the longer term the decrease from that initial error rate could be reduced substantially.

Residual problem statement

The effect of the residual problem statement, was visible and positive in almost all combinations of parameter settings.

The FFNN's, as mentioned before, were struggling with capacity problems. Therefore they could benefit from the simplification. The combinations with at least the raw values and stage one as data and fitting the residual problem clearly performed better than all the other combinations. In this situation, having the additional bands of last week's values did not contribute and could be omitted with another gain in accuracy.

For the LSTM models, changing from the traditional to the residual problem statement led to better predictions for all sensible combinations of the other variables. In the illustration of the results, two pairs of combinations clearly stand out: The one using raw data and the results from stage one (the two brown lines) and the pair using all three groups of variables (purple and turquoise). In both cases, the very important short term prediction could be reduced by about one, which clearly stands out. For predictions further out the effect was much larger in the first pair compared to the second, but even in the second, the effect was much stronger than the random uncertainty.

5.7 Conclusion and outlook

Proposed solution

This case study proposes two building blocks for use in deep learning based short-term traffic predictions than can be used regardless of architecture: The first building block consists of fitting a simple and robust estimator for the values in the form of the median over several of the preceding weeks. The second building block consists of fitting not the actual flows themselves but the residuals from the simple and robust estimator from building block one. While the second block is contingent on the first, the first can be used in isolation, in addition to or replacing the preceding week's value as a feature for prediction.

Effect on prediction

Those two building blocks were tested on two popular architectures for traffic flow prediction: Feed-forward networks and recurrent networks using LSTM cells. For both architectures the combination of the two building blocks led to increases in prediction accuracy that were most pronounced for somewhat wider prediction horizons. In the feed-forward networks, this was achieved by replacing the preceding week's values with the robust estimation and fitting the residuals, in the LSTM based networks, the results of the first stage had to be used in addition to the preceding week's values to achieve the best results. For both tested architectures, using both building blocks simultaneously had the additional effect of even decreasing the error rate for the very short term predictions in addition to the reduction of the increase of the error over the time of the prediction.

Extension

The architectures for which the effectiveness of the proposed changes have been shown have the limitation of scaling poorly with an increase in the number of stations under study. It remains to be evaluated whether the proposed changes to the way prediction is performed also have a benefit on architectures that scale well, e.g. graph based architectures. Alternatives could include attempts at dimension reduction using for example a PCA (Elhenawy and Rakha 2016) or graph networks (Shahsavari and Abbeel 2015). As the building blocks are not dependent on the architecture there should not be any obstacle to adding them to graph networks or any other kind of architecture.

Chapter 6

Discussion

The difference between perseverance and obstinacy is that one comes from a strong will, and the other from a strong won't.
— Henry Ward Beecher

This thesis provides a contribution to three steps needed to obtain insights from individual, passively tracked trajectories. They consist of establishing the individual geometry from sparse and potentially inaccurate data, enriching this data semantically, and using aggregates to gain actionable information.

All of this is set against the backdrop of two large developments. First, the ever more established big data technologies that allow the storage of the large quantities of data required to perform these steps on a population scale at high temporal granularity at telecommunication companies (Jony et al. 2015). Second there is the evolution of the mobile phone standards that provide ever increasing spatial accuracy (Chaloupka 2017; Peral-Rosado et al. 2012).

Although ample methodology for those steps was already in place, for the most part that methodology relied on the properties of existing sensors. Methods to semantically enrich trajectories often rely on the spatial accuracy of GNSS and traffic flow prediction methodology for the most part relies on fixed sensor grids built into the street network. To explore whether those same methods can be applied to the properties of the new data sources of passively tracked mobile phone data and how this can be improved in light of the limitations of that source were the aims of this thesis.

This section will place the contributions made and the insights gained in the earlier chapters of this thesis against the overall research aims presented in Section 2.4. It will discuss in detail the extent to which those aims were achieved, what insights were gained, and where the limits of the proposed solutions were identified.

A final section will treat possible implications of the work presented in this thesis with a focus on data privacy.

6.1 Reconstructing geometry from CDR

The aim of Chapter 3 was to address the research gap identified in Section 2.4.1. In particular, using CDR's, a form of passive tracking already used in various domains (Amini et al. 2014; Steenbruggen, Tranos and Nijkamp 2015), a way of recovering the geometry of the movement of individuals was to be developed. This was to be achieved without making too many a priori assumptions, lest the advantage of passive tracking – its potentially global scope – could be lost.

6.1.1 Insights

Sparsity of the data

Using CDR's as a starting point for the reconstruction of the geometry of movement was motivated by the fact that they are already collected today on a large scale by many telephone companies. While this promises a relatively wide applicability of the findings obtained by the study, it comes at the increasing disadvantage of temporal sparsity as internet based forms of communication decrease the necessity for text messages and traditional phone calls.

Indeed, in the data set used to answer the question, sparsity was a prominent feature: Starting from an already moderate number of between two and three CDR's a day, a significant fraction of the CDR's was immediately followed by another, as demonstrated in Figure 3.2 and already observed in the literature in different areas of human behaviour (Barabasi 2005). This makes the effective temporal sparsity of the CDR's even more of a concern and emphasises the need for methods that work with relatively little information on any given day.

In the simulation study, it was shown that even for those users who were simulated with at least 3 CDR's a day on average, the proposed algorithms could achieve results comparable to those from the real life data set, with the proposed DAMOCLES method slightly outperforming the transaction based one. For users with almost no CDR's on the other hand, DAMOCLES simply could not cluster days sensibly and therefore could not really improve on the other approaches.

Passive tracking captures every person carrying a cell phone. In order to benefit most from this global scope, methods for the analysis of those data should exclude as few users as possible, because this exclusion can incur bias in the results, if the property of interest is correlated with cell phone usage. One cause for exclusion, low numbers of CDR's, could be mitigated by moving from trajectory based approaches requiring eight Jiang, Ferreira and Gonzalez (2017) or twelve (ibid.) CDR's a day to methods that interpolate using prototype days, reducing the necessary number to three.

Distance between days

The core assumption on human behaviour that underlies (part of) the methods presented in this case study is that there are days in the lives of people that for the most part look the same, geometrically, which seems reasonable given the high overall predictability of human movement found by Song et al. (2010) and the dominance of small topological motifs of daily movement identified by Schneider et al. (2013).

However, assumption does not necessarily need to be true. Some users exhibited behaviour that was very far from that expected, showing neither a dominant daytime nor night-time location and visiting very few places regularly and at irregular times. In such cases DAMOCLES cannot and in fact did not work properly, because not enough extended days were sufficiently close to cluster them together into a prototype day. As the users included in this study were in no way representative of a wider population, the precise extent of this problem cannot be estimated based on the available data.

For users that exhibited clearly discernible patterns in the visualisation of the GNSS data, the distance measure between days was able to represent them, as illustrated by the distance matrices.

Reconstruction error

The reconstructions obtained through both the DAMOCLES and the transaction based reconstruction scheme were major improvements over simple reconstructions based on a priori assumptions such as repetitive days or repetitive workdays, which are common in direct extraction of semantics from CDR's (Ahas et al. 2010; Calabrese et al. 2013). The median reconstruction error could thus be reduced by about 30% from 1480m to 1045m and the percentage of days with an average distance below 3km could be increased from 61% to 89%. Qualitatively, these improvements turned out to be true even for users who conform most strongly with those assumptions, as evidenced in Figure 3.16.

For 86% of observed users, there were at least 10% of days that were not captured well with CDR's, resulting in average distances above 3km, as shown in Figure 3.17.

While this means that both the proposed schemes for geometry reconstruction can indeed be used to obtain a better estimate of the whereabouts of people than schemes based on simple assumptions, the median daily error is still in the order of kilometres, which is not sufficient for many approaches to extracting semantic information from geometries.

However, part of this rather large uncertainty can be attributed to the fact that the reconstruction happens at the level of the cell location. Using the signalling data at ground truth temporal granularity even knowing the cell ID perfectly at every point in time leads to a median average reconstruction error of almost 600m, with still 37% of users having at least 10% of their days with reconstruction errors above 3km. A certain limitation is thus inherent in the data source. Better passive localisations from cell phones are possible today (Widhalm et al. 2015), although are not as common as CDR's in research.

6.1.2 Contributions in context

The two main methodological contributions were the following:

- *Representation of space*: In the literature on CDR's cells or a group of neighbouring cells in the Voronoi tessellation are often used to obtain important locations. GNSS based models, whose data is continuous to begin with achieve the discretisation of space by using buffers whose size is chosen to match the error rate of the position estimates. The custom of identifying some sort of cluster of important places of a person was adapted to the space of mobile phone cells. To achieve this, an analysis of the relationship between GNSS locations and cell ID's was required (cf. Figure 3.11), which was enabled by a dataset containing both types of information. The insights from this analysis allowed the rescaling of the distance matrix between cell positions and thus a clustering that was meaningful in both densely and sparsely populated regions.

- *DAMOCLES*: A pipeline (cf. Figure 3.12) for reconstructing mobility geometries based on the warped space described above was developed. This included a way of extracting prototype days from sparse CDR data and using those prototypes in the reconstruction of the gappy position vectors representing the sparse information available from CDR. It relies on a distance measure between two days that captures both temporal and spatial aspects of the daily representations used. In particular, non-populated entries of the vector-representation had to be handled with care to obtain meaningful clusters.

Additionally, a method formerly used with GNSS based prediction on a topological representation of movement was adapted and used to reconstruct empty parts of the daily vectors: a simulation study was then conducted and the results were compared to benchmarks that rely heavily on a priori assumptions.

The methods presented in this case study were able to work with lower numbers of CDR's when reconstructing the geometries of people's movement than what is found in the literature. This can help counteract the decrease in CDR's generated as a result of the increase in internet based communication. The limitations on spatial accuracy both inherent in the data and by design in the methods mean that the applications will for the most part be restricted to improving estimations of time-varying population densities in regions where due to CDR sparsity, counting CDR's per cell would bias the results.

6.1.3 Limitations and extensions

Limitations on accuracy

The biggest limitation to achievable accuracy, as discussed, is the error introduced by using information on the cell as proxy for the actual position. This cannot be overcome and is a fundamental problem with CDR data.

While accepting this limitation was chosen for the purposes of this study, two other ways are also possible. First, one could attempt to make assumptions to bypass the limitation. For stay segments at frequent location this could mean using the longitudinal nature of such data to improve the estimate of the position, using for example the frequencies with which CDR's were generated through different masts close to the location. Using some sort of weighted average instead of the normal mean to obtain the position correctly could prove to be a better estimate. Furthermore, and if one is willing to incorporate domain knowledge, one could of course try to "guess" which was a "likely" actual position. This approach would potentially use some GIS data to inform the estimations, which was the approach pursued by Renso et al. (2013).

A second way of dealing with the limitation is by restricting the scope of analysis so that the limitation is not too much of a problem in the first place. Assuming that short movement is not resolvable given the data, analysing only those movements that are somewhat longer and can therefore be captured at the cell level would solve the problem. This would require some reasoning about the upper and lower bounds for stay and move segments to infer semantics, as shown in Widhalm et al. (2015).

In addition to the cell granularity, there is another source of errors in the spatial accuracy: The goal was to reconstruct the actual movement of the tracked people. This is of course not possible, strictly speaking, as even a GNSS signal does not capture this due to its inaccuracies and finite sampling rate. For the purpose of this study the GNSS data recorded by the cell phones was treated as ground truth and the point of reference against which deviations were measured.

Any attempts at capturing the error incurred by GNSS data would require sensors that capture movement with more accuracy than what GNSS provide and with fewer gaps, which remains a problem inside vehicles. While this could be resolved in very dense mobile phone networks in the future (Dammann, Raulefs and Zhang 2015), the option that comes to mind for now are video recordings using advanced computer vision techniques (Zeisl, Sattler and Pollefeys 2015; Ziegler et al. 2014), which of course would again limit the scope of the data collection.

Vector-representation of days

The last limitation to be discussed results from the representation of the prototype day. The *extended days* (cf. Section 3.3.2) are vectors of fixed length, enforcing a coarse temporal granularity. This of course limits the best achievable accuracy of the reconstruction, as the changes in position can not be resolved finer than the granularity of the extended day. Differing from the limitation above, this limitation is a consequence of the method and thus was imposed by design.

The cost of abandoning it would be a new distance function between days that no longer depends on the time slots in the way DAMOCLES does. Such a new cost function would still have to incorporate the temporal distance (on the daily cycle), and would in addition have to be able to deal with sets of CDR's for comparison. Those sets would most likely be necessary, as bursts of CDR's can originate from different masts, all of which can contain information.

The benefit of working with continuous time is twofold. First, If a routinely performed activity ends half way through a slot, being able to reflect this in prototype days would reduce the errors, as the reconstruction would be free to assign the remaining half to whatever activity comes next. Second, such flexible schemes could potentially include information offered by bursts of CDR's. While on movement, bursts of CDR's can reveal this dislocation (if the last CDR was registered "far" from the first). Breaking up the slots would allow for a representation of such movement which can be expected to be closer to the true geometry. In DAMOCLES, currently at most one of the CDR's gets used (if the entire burst is within a time slot), erasing all information about direction and precise timing of the movement.

6.2 Mode detection from CSD like data

The research gap identified in Section 2.4.2 asks how close passively tracked positioning information needs to be to GNSS signals in order to allow for extraction of the semantic information of transportation mode.

Transportation mode detection based on GNSS data has been extensively studied (Shen and Stopher 2014), but as passively sensed data are not (yet) collected at the spatial precisions and temporal resolutions of GNSS data, it remains an open question as to how close the former must get to the latter for the established methods to work.

6.2.1 Insights

Spatial accuracy

Spatial accuracy was one of the two main factors that were analysed to determine the requirements for passively collected mobile phone information. Even for the largest spatial uncertainty, the results – in particular for the well detected modes – did not suffer that much, indicating that even at relatively large spatial uncertainties a classification is possible in principle. For the most stable combination of classification – segmentation and random forest without smoothing – the accuracy only dropped from 80.7% to 80.4%, kappa dropped from 0.64 to 0.62 and the edit distance increased from 0.68 to 0.69. However, there was no case in which using less distortion did not make results better. This is not entirely surprising given that less distorted trajectories reflect reality more closely.

The spatial uncertainty chosen was set to a value that roughly reflects the order of magnitude possible with passive tracking. Thus, the results mean that even with today's technology, from a technological point of view passive mode detection is a possibility, even without prompting the cell phones for a position update, which would also be possible at large scales but impose a burden on the mobile phone users in the form of reduced battery life.

Temporal granularity

The ideal temporal granularity depends on the spatial accuracy. In agreement with the literature, where the temporal question was asked in isolation (Bolbol et al. 2012), sampling rates of approximately one a minute seems adequate with higher rates only potentially being beneficial if the signal to noise ratio is high, i.e. if the spatial uncertainty is relatively low.

With high spatial uncertainty and the robust classification procedure including segmentation and random forests but not smoothing yielded accuracies from 77.8%, 80.4% and 75% kappa values of 0.66, 0.69 and 0.61 and edit distance values of 0.66, 0.62 and 0.63 for 5, 30 and 300 second intervals respectively. This demonstrates that with high spatial uncertainty too high of a sampling rate does not fare well with the used classification procedure.

Signalling data today is generated at a fairly high frequency, which is in the orders of magnitude used in the study. Collecting such data for all customers over extended periods of time requires big data storage which today is relatively easily available, removing the technical difficulties of handling temporal granularity as a limiting factor and paving the way for mode detection on a large scale.

It is notable that even the lowest temporal resolution used in this case study is significantly higher than what was available for the CDR study in Chapter 3. The effect of this difference is of qualitative nature and represented in Table 3.2: CDR data cannot be viewed as some sort of downsampled trajectory and have to be treated differently, for example by using DAMOCLES. The data used in this sensitivity analysis, while still rather sparse at the lowest level of granularity, could still be treated as trajectory.

Context in classification

Throughout the study, it became clear on various occasions how important context is for classification of transportation modes. Context has been found to help the analysis of movement in a variety of ways. Bleisch et al. (2014) have found different movement behaviour of fish in the Murray river depending on environmental conditions and Knoblauch, Pietrucha and Nitzburg (1996) have found pedestrians exhibiting different velocities depending on context such as the weather. In this studies, several different ways to capture context were employed.

First, there is the temporal context. The evolution of features over time helps understanding the underlying movement. This was shown by the fact that pointwise classification was by far the worst among the tested procedures: the points lack temporal context. However, the ways that incorporate this context, be it with a posteriori smoothing, classifying the sequence instead of points, or identifying segments that could be expected to be the same mode, all improved classification results significantly. In this study, the optimal variant was to segment the trajectory and use window based features. Comparisons between fundamentally different approaches to incorporate temporal context are sparse in the literature, and it is impossible to tell, whether the impressive results for example from Stenneth et al. (2011) could be improved by adding a segmentation step.

Second, there is spatial context. In the study the effect of spatial context on the classification results were investigated, in accordance with the majority of the literature, even if some work has been conducted using spatial context in segmentation (Liao, Fox and Kautz 2007). Having GIS information was beneficial to classification. For reasons explained in Chapter 4, this effect was small in the study presented in this thesis but can be assumed to be underestimated. Especially the modes that can profit most from GIS information – buses and trams – had very high feature importance values for the corresponding layers of GIS information reflecting that importance. The literature suggests that the more specific GIS information is, the more it contributes to classification, as Stenneth et al. (2011) showed.

Finally there is the semantic context. It was surprising to see how well the mode *train* was detected in the case study, especially noting the low feature importance of train related GIS features. However, as for this mode the important features revolved around distances between points. Given the sparsity of points for this particular transportation mode and the frequent gaps due to signal loss, it stands to reason that the way the mode of transportation influenced the data collection had an influence on classification. These effects are hard to plan for, do usually not take the form of specific features, and are not commonly discussed in the literature. One should nonetheless be aware of them when evaluating the results of a classification. After all, they could, as in the case in the case study in this thesis, have a positive impact on classification that can impact generalisability of the results. In the transportation literature, the effect of the semantic level on data collection has so far not been extensively studied, except for the case of gaps in subways, where accelerometers are used to fill the gap Widhalm et al. (2012).

All of this goes to show that although superficially transportation mode detection looks like a textbook example of classification for which textbook classifiers can be expected to perform admirably, spending time thinking about what the signal *means* and incorporating this into the classification will improve the results. This may seem trivial, but all those little idiosyncrasies of transportation, such as the mismatch between characteristic time scales of mode changes and collection rate, GPS signal losses that are correlated to transportation modes, and the likelihood of certain combinations of transportation modes are all there to be made use of for getting a more accurate image of how the people under study moved about.

Cross-validation

The sensitivity analysis showed that the way in which cross-validation is performed can play a role in the outcome and insights obtained. This deserves mentioning explicitly here as fairly often in the literature there is no explicit treatment of how cross-validation or the split into training and validation data is performed and sometimes when it is stated, it is indicative of overly optimistic results (Feng and Timmermans 2016). As was shown in this case study, cross-validating randomly will lead to more optimistic results, and to greatly reduced variance in the generalisation error. The importance of the independence between training and validation data found in textbooks such as Hastie, Tibshirani and Friedman (2017) is thus once again be clearly demonstrated.

Label quality and segmentation

In the collection of the trajectories, the stages of the trips were identified by the system and users were not able to introduce additional ones, leading to many missing access/egress stages in what was used as ground truth data. This led to the bad performance of those transportation modes that could make use of GIS information on (bus and tram) stops, since the connection between those stops and the beginning or end of a corresponding stage was obfuscated. However, overall statements of the study can still be made, as the aim was not to achieve state of the art results but to investigate the deterioration of the results incurred by decreasing data quality.

The segmentation attempted to capture all potential points where the transportation mode could have changed by identifying the stopping points. In light of missing access/egress stages this is bound to produce mismatches between segments and stages, yielding overly pessimistic results. Eliminating that mismatch completely by using the stages as segments can only give a partial image of the separability of the transportation mode as it also eliminates part of the problem to be solved. The marked increase in classification quality observed nonetheless demonstrates the importance of a high quality segmentation and the potential for smarter ways to perform it.

The common way to collect data is to work with just a few people that are well instructed and may even have a personal interest in its quality (Feng and Timmermans 2016; Stenneth et al. 2011). This can be conducive to data quality, but comes at the cost of variability and thus therefore can lead to overly optimistic results.

6.2.2 Contributions in context

The study presented a comprehensive analysis of the sensitivity of commonly used transportation mode detection techniques to spatial and temporal uncertainty. While the temporal question is also of relevance in a GNSS context and has – in that context – already been investigated by Bolbol and Cheng (2010), the application to passive tracking necessitated a combined analysis of spatial accuracy and temporal granularity which this study provided. It revealed an interaction between the two dimensions that should be heeded when conducting mode of transportation detection based on spatially less accurate data. Furthermore, the study showed that with the passive positioning and the storage capability available today telephone companies could perform large scale mode detection, if they invest in collecting high quality labels.

A very diverse and quite sizeable group of people was used in this study to provide the data. As a result, the data was spread over almost the entire country, thus capturing a multitude of mobility experiences from remote villages in the Valais to the urban hustle in Zurich. Inevitably, most of the data originated in big cities, reflecting the population distribution in Switzerland. This method of tracking people is in contrast to common settings. The number of participants is often limited: six in Stenneth et al. (2011), eight in Feng and Timmermans (2016), fourteen in Nitsche et al. (2012) and sixteen in Reddy et al. (2010). Furthermore, often staff members of research facilities and their friends and families collect the data (Gong et al. 2012; Kiukkonen et al. 2010; Nitsche et al. 2012), if the recruitment is mentioned at all. Both those decisions will lead to spatially more homogeneous data, potentially facilitating the classification task.

The diversity of approaches used for classification was another distinguishing feature of the study. Although many studies compare different classifiers, they typically refrain from combining entirely different approaches to classification, such as pointwise vs segmentwise classification, or forms in between, such as CRF. While the study does not overcome the comparability problem of transportation mode detection, it at least allows a broad comparison of different approaches on the same dataset.

The broad comparison of different approaches in this case study revealed that transportation mode detection is possible based on data in the accuracy that is available today from passively sensed data.

The spatial diversity of the participants together with a careful cross-validation ensured that the results are not overly optimistic, and the breadth of approaches taken for classification contributed to the results not just being a result of a specific chosen approach.

6.2.3 Limitations and extensions

Altering ground truth

As mentioned previously, the fact that users could not add or remove stages posed a problem, as what was used as "ground truth" turned out not to be so, as illustrated by the mode sequences that were observed.

One possible way to approach this problem could be to alter the labels so that they *make more sense*. This could ensure for example that every trip starts and ends with a walk segment, force intermittent walking segments between public transport modes, and so on. Additionally, the points where people change transportation mode to and from public transportation could be moved to appropriate stops.

This would undoubtedly increase classification performance as it would ensure the significance of GIS features (because they were used altering the labels and geometries) and the completeness of access/egress stages. The precise magnitude of this increase cannot be determined based on the case study alone and would require further investigation. However, altering the ground truth in this way is not unproblematic. Pre-processing in general does not pose a major issue, as long as it can also be done on the testing data where the labels are unknown. It is the fact that the labels are used to determine where an additional stage should be and where exactly the breaks between stages are that is problematic.

This step would therefore present an upper bound of what could be achieved with the data at hand. To make a realistic estimation about the highest achievable classification results on this realistic dataset, the process would have to enforce strict boundaries between what can be used in training and what cannot.

Segmentation

There is a significant increase in classification quality if, instead of the calculated segments, the stages are used as segments. This does not come as a surprise as it reduces the combined segmentation and classification problem to one of classification alone. Nevertheless, the size of this increase poses the question as to how much better the results could be if the segmentation was not merely based on stops.

The fundamental difference between segments (at least the way they were calculated in this study) and the stages lies in the fact that the segments are purely geometric, while the stages are semantic. One avenue that could be pursued in this direction would be a move towards "semantic segments". This is not to be confused with the concept of semantic trajectories as used by Parent et al. (2013), where the semantics are added to extracted geometric segments. Instead, the segments are inferred having the application domain in mind.

The idea would be to look for other features beside speed that, while geometric in nature, are very closely related to semantics. The fact that they would likely be used in addition to the speed based segmentation would most probably necessitate a generalisation of what segment based classification means, because the different layers of semantics would lead to different but simultaneous segmentations. For example, it is possible to imagine a segmentation based on bus proximity, another based on presence on roads that are impassable for cars, and so forth. All these segmentation layers would then have to be incorporated into a generalised segmentation.

An alternative, perhaps somewhat less radical way of changing how segmentation is performed is to abandon the realm of rule based systems and try to learn the segmentation as well as the labelling. This would have to be done in a way that ensures the segments resemble the stages and with larger penalties for under-segmentation than for over-segmentation, due to the relative ease of combining stages with the same label. One could approach this in multiple ways. One could start, for example with a pointwise classification into *segment break* and *no segment break*, followed by some smoothing to unify multiple breaks in close temporal proximity. More involved schemes could look at the whole trajectory and return a set of break points in the interval $(0, 1)$ using for example attention based methods.

Classifiers

The classifiers used here, while common in the literature (cf. Section 2.2), do not represent the current pinnacle of machine learning. As a machine learning problem, transportation mode detection is one of sequence labelling. This makes it conceptually similar to speech recognition, where the time interval between two adjacent measurements of a microphone (corresponding to the fixes in mobility) is significantly shorter than the duration of a phoneme (corresponding to a stage in the analysis of movement) (Trigeorgis et al. 2016).

Consequently, advances in the field of speech tagging could be translated to the domain of transportation mode detection which so far has seen relatively little impact from anything related to deep learning. With deep learning, there is always the question of whether enough data are available. However, because the number of transportation modes is small and their sequences are simple compared to their equivalents in speech (dozens for parts of speech (Plank, Søgaard and Goldberg 2016) and dozens of phonemes (Sejnowski and Rosenberg 1987; Weide 2005)), the need in terms of data size could be in the order of magnitude that can effectively be collected, although this would have to be investigated.

The requirements of such a system would be that it can handle any length of data, as trips range anywhere between five minutes and several hours. However, as this wide range of different sequence lengths is also a reality in language, solutions are bound to be available. Furthermore, the solution should encourage short sequences of output labels, thus shrinking the input of potentially hundreds of input vectors to only very few output labels; again, however, there are solutions that accommodate this kind of requirement.

Even sampling rate and error distribution

One possible limitation of the study is that of the evenness of the sampling rate. While it is not exactly even because of signal loss, there are certainly no bursts that could be present in passive cell phone data.

One effect potentially present in passively tracked data but not in the data used for this study could be the combination of varied spatial uncertainty and varied temporal granularity. While the two were held steady for every analysis, the effect of a altering them within a single trip cannot be observed in the setup of this case study.

One way this limitation could be addressed is by assuming a stochastic process for the time between fixes and sampling according to this process. The challenges to be overcome with this approach are, first, to obtain a reasonable estimation of how such a process should look from actual CSD data. Second, as the actual sampling that was available in the data used for the case study was (mostly) in the order of seconds, the mismatch between a realisation of the stochastic process and the acutal times at which data is available would somehow have to be bridged.

A second limitation concerns the assumed spatial error process of the data. Again, it would make sense to study the process of combined GNSS and passive mobile phone tracking positioning in depth and model it thereafter. However, this solution is dependent on data that is still very hard to obtain in practice.

Variable selection and sample size

The final limitation discussed here is that of variable selection and the fairness of comparison between different classification schemes. In the study the same set of variables was used for all classifiers. However, while some classifiers are well suited for large numbers of features, such as the random forest, others rely more heavily on not having too many noise variables. The question therefore arises as to whether the comparison as presented in the case study was fair.

One solution would be to perform a variable selection through a stepwise process for all of them, choose a criterion by which to decide on the optimal set of features and report those results. Given the rather large set of variables used, this would incur a significant increase in the work load.

Similarly, some classifiers do not scale well to large sample sizes, such as the support vector machine. Once again this raises the question as to whether it is fair to compare the algorithms based on different sample sizes. While the answer seems to be no, in this study, SVM was trained only on a subsample of the training data. This was chosen as training it on the entire sample would simply not have been practicable due to a runtime which is above quadratic in the sample size. The aim of mode detection is to apply it to real world datasets, which means that it should be possible to transfer and actually use the identified solutions. Therefore the comparison based on unequal sample sizes, which can be deemed unfair, was chosen.

6.3 Predicting single mode traffic flows

For passively tracked data there will not be any multi-year history of data available when the recording first starts and the mobility patterns may change faster than those observed on the stationary sensor network on freeways, as the possible sources for changes is larger due to the increased scope. Therefore, the research gap identified in Section 2.4.3 focused on incorporating domain knowledge in the features and re-formulating the traffic flow prediction problem in a way that makes it simpler to solve. This should improve deep learning based approaches that typically rely on large training datasets. The first part was achieved by first using a simple but robust model for the hebdomadal cycle instead of simply using the previous week's values, reducing the impact of special events. The second part was achieved by restating the traffic flow prediction in its residual form, taking some of the complexity out of the prediction.

6.3.1 Insights

Domain knowledge should inform architecture

Much of the literature focuses on comparing different architectures that fared well in other domains (Shao and Soong 2016; Wu and Tan 2016). The introduction of new and admittedly sometimes very powerful ways of setting up the deep network is unquestionably important and can help to improve the prediction of traffic flows.

However, this case study has shown that architecture itself is not everything. Echoing the findings from the transportation mode detection case study, context matters also for traffic flow prediction: The traffic flows happen inside a well studied environment and there is no reason why this knowledge should not also be exploited for prediction. Both adding the simple robust estimator and restating the prediction problem in a residual fashion could reduce prediction errors in both tested architectures. These findings are parallel to those in other domains such as remote sensing, where for instance incorporating the fact that the rotation of the image is irrelevant contributed to simpler and thus smaller models (Marcos et al. 2018).

This is not to say that architecture is not important, as became clear when comparing the results of the feed-forward and the recurrent networks. The recurrent architecture outperformed traditional feed-forward architectures and so if in practice not both architectures can be tested, using recurrent architectures seems preferable.

Even in cases where the architecture is not optimal, the two proposed solutions could therefore help to achieve better results. The feed-forward networks had capacity problems, but the nature of the data limited their potential to fit very large models. By making the problem simpler and by increasing the expressiveness of features, some of the limitations of the architecture could be mitigated.

Importance of the flow itself

It was surprising to see that, contrary to the mode detection case study where adding context variables could improve labelling quality, adding thematic information in the traffic flow prediction case study only improved predictions for the FFNN network and, even then, only helped for predictions three hours or more into the future (cf. Figure 5.4a). Instead, it appears that the traffic flows already contained all the necessary information for prediction, and if the architecture allows to use it, as was the case for LSTM based networks, context variables could not contribute to the prediction.

6.3.2 Contributions

Robust estimator and residual problem statement

The primary contributions of this study are the robust estimator as a first stage of prediction and the residual problem statement, both of which are new in traffic flow prediction based on deep learning.

The first contribution is an attempt at improving on the common practice of providing information about the preceding day or week as covariates. The robust estimator used instead is more stable and therefore a more reliable feature to use in predictions, as evidenced by the results (cf. Figure 5.5). It trades in some increase in bias incurred by the larger temporal window of the estimation of the covariate against a reduced variance achieved by considering multiple points in time that are known to have similar characteristics. As became clear from the results, the benefits of this trade-off outweigh its costs.

The second contribution is an attempt at removing the seasonality from the signal in order to make the problem simpler for the estimator. This idea is widely known in time series analysis and has been proven to work in deep neural networks but had not yet found its way into deep network based traffic prediction.

Error definition

A secondary contribution was the use of a prediction horizon that was not limited to a single point in time. Traditionally traffic flow prediction is concerned with the prediction of a fixed time horizon, very often one step of one to fifteen minutes ahead (Vlahogianni, Karlaftis and Golias 2014) and the results on further prediction are either not reported, or reported only after another model has been fitted to work on this new time horizon (Lv et al. 2014).

Optimising over a whole range of time intervals allows for an evaluation of the quality of traffic flow prediction over a time horizon that matches the orders of magnitude relevant for decision making, as the trips alone can take hours (Kung et al. 2014).

Explicit estimation of random error

All results obtained through deep learning are essentially based on the outcome of a pseudo random number generator. In this case study the effect of this randomness was stated explicitly by reporting the ranges of the results over multiple runs. Thus it can be compared to the effect of the different choices in the models throughout Section 5.5. This increases the confidence in the actual effects observed resulting from the proposed changes to traffic flow prediction.

Although uncommon in the literature, it would be a simple measure to improve the transparency of the results as they can be put into perspective. It is clear that this multiplies the computational effort required to calculate all the results, but the added transparency definitely makes this worthwhile.

Both innovations to deep learning based traffic flow prediction, the robust estimator and the residual problem statement could reduce the prediction error, particularly for but not limited to somewhat wider prediction horizons. Especially if used in combination, the deterioration of the prediction accuracy observed for larger time horizons could be significantly reduced.

6.3.3 Limitations and extensions

Tested architectures and system wide prediction

The aim of the case study was to demonstrate the importance of factors besides different architectures. As such, it compared only specimen of the two dominant architectures found in the literature. One of the types absent in the case study was the graph based architecture (Shahsavari and Abbeel 2015).

The reason for this was that the graph based architecture inherently works on the entire graph, whereas the commonly used architectures run into performance problems if one attempts to fit them on the thousands of sensor nodes in the entire system under study. This would result in problems of comparability so that the significant additional effort to build the second pipeline could not be justified. Knowing the effect of the proposed improvements on graph based convolutional architectures would of course be valuable and it can only be hoped that it will be investigated in future work.

As well as testing another type of architecture found in the literature, this would overcome a second limitation of this study: its limited scope in terms of the stations used. This limitation haunts almost the entire literature on short-term traffic predictions, but of course overcoming it would bring a comprehensive prediction of the traffic situation of an entire region. This in turn would enable a wide range of applications for which today a large number of separate models would be required.

A second type of architecture that was not included are the naïve convolutions applied on either parts of the network with a very simple topology of a straight line, (Wu and Tan 2016) or the rasterised information on the network stripped of its topology (Yu et al. 2017). The reason for this is that both types significantly simplify how they represent their information and thus limit their applicability already at the outset. Standard convolutions on non-forking parts of the street network cannot easily be extended to cover larger parts and will therefore remain limited in their application. Rasterising the traffic information will incur significant information loss as for example the two directions of a street can no longer be separated.

Covariates

The study considered only temporal and meteorological covariates. This entails at least two problems. First, in the area under investigation, the impact of the weather may be limited by the importance of the motorised individual transportation. Essentially, there are only few alternatives irrespective of the weather conditions. Thus, the impact of the weather is limited to altering the number of trips that are undertaken. In other regions, the weather can also have an impact on mode choice, thus giving it higher leverage over the number of vehicles on highways. Furthermore, the weather in Los Angeles is not as strongly varying as in other parts of the world, limiting its impact even further. Therefore the small magnitude of the effect of the weather can partly be attributed to the choice of the study area and may be different elsewhere.

The second caveat of the choice of covariates in this case study is the fact that other factors – such as social events – were neglected. In their absence, the only way for the model to realise the presence of a special event is through the traffic itself. However, by the time this is detected, of course the prediction has already been wrong. There are several ways in which such information has been incorporated into models, for example using the approach of Chen et al. (2016). Although it would have been an option to include it in the case study, finally this option was discarded as being too far removed from the core message.

6.4 Privacy impacts

As presented in Section 2.1.4, this thesis was able to avoid any problems with data privacy by working with data that either was not personal data in the first place – the Eulerian highway loop detector data – or where the data subjects gave explicit consent for their data to be used for research purposes. However, if the results of this thesis find their way into operational applications, the situation may be different. This section therefore discusses possible implications of this research when applied to data that was not collected for the purpose of research.

The important distinction to make here is the privacy in the input data versus the privacy in the results of the analyses. If only the latter has to be guaranteed then even very sensitive data can be used in the analysis. If the former is also relevant, then the privacy of the data subjects are even protected from the analyst, reducing the potential for privacy breaches. As discussed in Section 2.1.4, privacy is often achieved by summarising information in a way that avoids individuals being identified at the cost of a reduced usefulness for further analyses. As the results of the analyses presented in this thesis can always be aggregated in a privacy preserving way, the focus of this section lies on privacy preserving data as input.

The reason why geospatial information is so sensitive is the combination of their power to identify people (De Montjoye et al. 2013) with their semantic expressiveness (Rinzivillo et al. 2014). Both those aspects can be diminished by sacrificing input data quality either spatially – such as by using data on the cell tower level – or temporally by using reduced sampling rates, both of which have been done in this thesis (Chapters 3 and 4) and would also be an option in applications.

The relevant question then becomes whether bad data quality is viable to protect the privacy of data subjects. One sufficient way of achieving this would be to render the data impersonal. However, simply reducing the quality can not be used as a measure to guarantee privacy. While a spatially coarse and potentially sparse data collection reduces identifiability, Culnane, Rubinstein and Teague (2017) argue that a scheme preserving privacy on the input side needs to break the temporal links between information about the same individual, even for short durations. As long as the links persist, the data thus cannot be considered fully privacy preserving. Some of the methods presented in the thesis rely explicitly on having longitudinal data on the data subjects (Chapter 3 and parts of Chapter 4) and are therefore not suited for privacy preserving analyses.

One exception to this is the transportation mode detection on features excluding longitudinal information. For example, a scheme where only anonymised trips – or even just parts of trips, such as segments – are available to the analyst, could be a possible data source, as indicated by the good performance of classifiers on segments in Chapter 4.

The other exception is the Eulerised data from the loop detectors (Chapter 5). For the purpose of flow/speed predictions, situations with few or no observations have the same meaning (free flow) and therefore any of the typical approaches to dealing with the problem of small counts can be applied to ensure privacy.

Taken together, these two exceptions could lead to a pipeline in which streamed flows of passive tracking information is cut into trips or segments first. These trips would already be far less problematic from a privacy point of view. In a second step, the trips or segments could be classified and the results fully rendered impersonal through Eulerisation and related steps. This Eulerised information could then be used for all sorts of analyses, such as flow prediction.

As this pipeline needs personal data to ensure the cuts between trips or segments are at the proper spatio-temporal coordinates it does not fully rely on impersonal data. However, as the links between data of the same individual get broken fairly early on in the process, it could be beneficial if there are multiple analyses that use the results of the first step as input. That way, irrespective of the number of data processors, only the first gets to see personal data.

Chapter 7

Conclusion and outlook

Victorious warriors win first and then go to war while defeated warriors go to war first and then seek to win.
— Sun Tzu

7.1 Revisiting research questions

Following the detailed discussion of the results of this thesis in Chapter 6, this section revisits the research questions and provides a distilled version of the contributions and what was learned.

7.1.1 Reconstructing geometry from CDR

The first research question, in the context of very sparse and inaccurate CDR's and how they should be used on a geometric level, was the following:

> **RESEARCH QUESTION 1:**
> How accurately can the movement geometry be extracted from call detail records using as few semantic assumptions as possible?

Quintessential answer

In terms of spatial accuracy, the reconstruction improved on the assumption of having a weekday-weekend routine by about 33% but remained in the order of magnitude of a kilometre. Further improvements on this are limited by the fact that cell ID's are inherently inaccurate: Using the best possible cell ID based reconstruction still has reconstruction errors just a few hundred metres better than that of the proposed methods.

Contributions

First it was established that CDR's were too sparse to be treated as a trajectory. This became evident from the analysis of summary statistics. In order to reason about movement, a representation of space was proposed that takes into account the different densities of mobile phone cells in urban and rural areas. Based on this representation of space, two methods were proposed for the reconstruction. The first method, DAMOCLES, clustered daily routines based on a distance measure that was designed specifically for this kind of data and used those clusters for the reconstruction. The other method was an application of association mining used in other domains and worked in atomised time devoid of temporal order.

In terms of the goal to avoid semantic assumptions, both proposed methods only assume that certain location sequences can be observed on multiple days. In the transaction based method, the repeated patterns are of an arbitrarily small length and could consist of information from just one slot. For DAMOCLES on the other hand, the repeated unit is the day, which is a somewhat stronger assumption. It has to be stressed that this does not imply that all days must look the same, and in fact an arbitrary number of prototype days can be discovered as long as there are enough of them to form a cluster.

7.1.2 Mode detection from CSD like data

Building on the insights into the limitations of cell ID based spatial accuracy, the second research question asks by how much the spatial accuracy would need to improve, and at what temporal granularity the data collection would need to happen for passive tracking to be able to perform an important task in traffic planning, transportation mode detection:

> **RESEARCH QUESTION 2:**
> How much worse than GNSS data can passively tracked data be in terms of spatial accuracy and temporal granularity while maintaining the distinguishability of transportation modes?

Quintessential answer

In general, transport mode detection is possible at the sparsest temporal granularity and the highest spatial uncertainty that were tested, reflecting estimates of what is possible today for passive tracking.

In terms of spatial accuracy, any reduction improves classification results and therefore techniques that help reduce it appear to be useful. This contrasts the situation in temporal granularity. If the spatial uncertainty is high, further refining temporal granularity will actually decrease the quality of the classification with the tested methods and a collection rate of around half a minute seems to be optimal.

Due to properties of the data, the validity of the findings are limited to the transportation modes of car, train and bike that are least affected by the bad label quality of access/egress stages. A separate analysis on segments that are based on the stages reveals similar effects of spatial accuracy and temporal granularity at significantly higher levels of classification accuracy. This can be taken as an indication that the findings can be generalised to mode detection in general.

The sensitivity analysis on cross-validation stressed the importance of the independence between the folds and revealed overly optimistic results in the case where this independence is not given.

Contributions

A careful sensitivity analysis of the transportation mode detection was conducted to establish the effect of temporal granularity and spatial accuracy of the data on the classification results. This sensitivity analysis was conducted on a broad dataset comprising over a hundred people from all over Switzerland, thus providing diversity in terms of the region under study, from urban Zurich to the rural Valais. In this dataset the transportation modes were skewed which had to be accounted for in the classification.

As the results from the literature are difficult to compare due to different datasets or different specifications of the mode detection problem, the analysis incorporated a wide range of possible solutions to the problem: In terms of the overall approach, both pointwise and segmentwise classifications were tested, with and without smoothing of the resulting labels. In terms of classifiers, a selection of five different classifiers that are commonly used in the literature were applied. In terms of features, different groups of features from GIS and based on different temporal aggregations were used.

Lastly a sensitivity analysis on the different choices was conducted. This included not only parameters of the feature construction and hyperparameters of the classifiers but also the important procedural choice of how to cross-validate such results.

7.1.3 Predicting single mode traffic flows

The feasibility of mode detection from passive tracking allows the analysis of traffic flows to be detached from dedicated, immobile sensors. Thus, questions that were hitherto bound to an immobile and expensive sensor network could be answered using passive tracking, warranting a look into the methodology commonly used in that field. The identified gap to fill was the following:

> RESEARCH QUESTION 3:
> How and by how much can the error in deep learning based traffic flow prediction be reduced by reframing the prediction problem and reducing its complexity?

Quintessential answer

For both tested architectures, applying the proposed approach to traffic flow prediction reduced the error rate over all tested time horizons, demonstrating the usefulness of the solution.

While the FFNN networks had capacity problems and therefore had to use the robust estimate as a replacement for the otherwise used values from the preceding week, the LSTM networks could make effective use of both the robust estimate and the preceding week's value.

The reductions achieved were 15% for LSTM and 8% for FFNN in the shortest possible time window (15 minutes) and 27% for LSTM and 11% for FFNN over the whole six hour prediction period.

Contributions

To answer the research question, the idea to simplify the problem that has been successfully applied to deep learning problems in other domains has been adapted to short-term traffic prediction.

The proposed simplification takes the form of a residual problem statement in which instead of the signal, the residual to a simple prediction is learned. The reason for this attempt lies in the well documented diurnal and hebdomadal patterns of traffic flow. If a prediction model does no longer have to learn those patterns because that signal has already been eliminated, then the model can put more emphasis on the deviations from the signal, leading to better results. While the flow values of the preceding weeks are often used as a predictor, explicitly fitting the residual constitutes a different problem to be solved.

The *simple model* that is used as the signal which is subtracted before the learning happens was chosen as the median over the last three weeks. This is an improvement over the typical use of the preceding week's value as it is more robust in weeks following special occasions. The variance could be further reduced at the cost of bias if the window over which the median is enlarged.

One of the advantages of the proposed approach to traffic flow prediction – residual problem statement based on a simple model – is it's independence from the type of architecture that is used and could be applied to any system used in that domain.

To demonstrate the effectiveness of the proposed solution, it was applied to two different architectures on the often used PeMS dataset of district seven.

7.2 Trends and outlook

In Chapter 6, the limitations of the individual methods used and choices taken in this thesis were discussed in detail. That chapter also outlined some limitations and possible ways of addressing them and thus provided avenues for future research in the specific areas covered by this thesis. This section therefore focuses on the overarching trends of the developments within the domain covered by this thesis that go beyond individual methods.

Passive tracking is a powerful tool, as is now being realised by large telephone network operators. In an attempt to compensate for decreasing revenues from traditional calls and text messages, network operators and service providers are searching for new sources of revenue and it seems inevitable that insights from passive tracking will be monetised. If successful, this will boost both the available data and the methodology used to obtain insights from the mobility of people.

This transition to ever more accurate positioning is assisted by developments in technology. Even with the developments of LTE that are already known today, accuracies rivalling or surpassing those from GNSS are foreseeable, even without the developments that can be the result of efforts to realise self driving vehicles.

However, all this high quality data will be collected by the network operators which is a mixed blessing for research. First, obtaining access to this kind of data is difficult. Being aware of the substitutability of their services, network operators are at least to date very hesitant about sharing that kind of information, as any breach in privacy could result in a PR disaster and a dramatic loss of customers. This is good news for privacy aware customers, but means that obtaining data may be harder for larger fractions of the population than it is for small sample sizes nowadays used in GNSS studies.

As building and maintaining the infrastructure for large scale tracking can be assumed to be expensive, a second difficulty with this kind of data will be that it will be hard to obtain it in the correct shape for the research in question. It can be assumed that researchers will have to accept the data that they have access to and conduct the research based on them, instead of tailoring the data collection to research needs, as is predominantly done today.

Despite all the caveats mentioned, large scale passive tracking is likely to become a very fruitful endeavour. On the one hand, there will be methodological innovation, as the very fact that large sections of the population are tracked allows for different methodologies, such as trajectory segmentations that not only depend on the geometry of the trajectories in question but also on that of *nearby* trajectories, that will become available. Thus the research fields of collective movement and transportation science could be brought together. On the other hand, there are insights on the semantic level that can be expected to be profound and to be looked forward to.

Bibliography

Primary references

Ahas, R., Aasa, A., Yuan, Y., Raubal, M., Smoreda, Z., Liu, Y., Ziemlicki, C., Tiru, M. and Zook, M. (2015). 'Everyday space–time geographies: using mobile phone-based sensor data to monitor urban activity in Harbin, Paris, and Tallinn'. In: *International Journal of Geographical Information Science* 8816.July, pp. 1–23.

Ahas, R., Aasa, A., Roose, A., Mark, Ü. and Silm, S. (2008a). 'Evaluating passive mobile positioning data for tourism surveys: An Estonian case study'. In: *Tourism Management* 29.3, pp. 469–486.

Ahas, R., Silm, S., Järv, O., Saluveer, E. and Tiru, M. (2010). 'Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones'. In: *Journal of Urban Technology* 17.1, pp. 3–27.

Ahas, R., Silm, S., Saluveer, E. and Järv, O. (2008b). 'Modelling Home and Work Locations of Populations Using Passive Mobile Positioning Data'. In: *Location Based Services and TeleCartography II: From Sensor Fusion to Context Models*. Ed. by G. Gartner and K. Rehrl. Chap. 18.

Ahmed, M. S. and Cook, A. R. (1979). 'Analysis of freeway traffic time-series data by using Box-Jenkins techniques'. In: *Transportation Research Record* 722.

Amini, A., Kung, K., Kang, C., Sobolevsky, S. and Ratti, C. (2014). 'The impact of social segregation on human mobility in developing and industrialized regions'. In: *EPJ Data Science* 3.1, pp. 1–20.

Anderson, R. L. (1970). 'Electromagnetic Loop Vehicle Detectors'. In: *IEEE Transactions on Vehicular Technology* 16.February, pp. 23–30.

Axhausen, K. W., Madre, J.-L., Polak, J. W. and Toint, P. (2003). *Capturing long-distance travel*. Research Studies Press.

Axhausen, K. W., Zimmermann, A., Schönfelder, S., Rindsfüser, G. and Haupt, T. (2002). 'Observing the rhythmns of daily life: A six week travel diary'. In: *Transportation* 29, pp. 95–124.

Balmer, M., Rieser, M., Meister, K., Charypar, D., Lefebvre, N., Nagel, K. and Axhausen, K. (2009). 'MATSim-T: Architecture and simulation times'. In: *Multi-agent systems for traffic and transportation engineering*, pp. 57–78.

Bantis, T. and Haworth, J. (2017). 'Who you are is how you travel: A framework for transportation mode detection using individual and environmental characteristics'. In: *Transportation Research Part C: Emerging Technologies* 80, pp. 286–309.

Barabasi, A.-L. (2005). 'The origin of bursts and heavy tails in human dynamics'. In: *Nature* 435.7039, pp. 207–211.
Batty, M. (2012). 'Smart cities, big data'. In: *Environment and Planning B: Planning and Design* 39, pp. 191–193.
Bayir, M. A., Demirbas, M. and Eagle, N. (2010). 'Mobility profiler: A framework for discovering mobility profiles of cell phone users'. In: *Pervasive and Mobile Computing* 6.4, pp. 435–454.
Bearman, A. and Dong, C. (2015). 'Human pose estimation and activity classification using convolutional neural networks'. In: *CS231n Course Project Reports, Stanford University*.
Becker, R., Cáceres, R., Hanson, K., Loh, J. M., Urbanek, S., Varshavsky, A. and Volinsky, C. (2011). 'A tale of one city: Using cellular network data for urban planning'. In: *IEEE Pervasive Computing* 10.4, pp. 18–26.
Becker, R. et al. (2013). 'Human mobility characterization from cellular network data'. In: *Communications of the ACM* 56.1, pp. 74–82.
Berndt, D. and Clifford, J. (1994). 'Using dynamic time warping to find patterns in time series'. In: *Workshop on Knowledge Knowledge Discovery in Databases* 398, pp. 359–370.
Bhat, C. R. (1998). 'Accommodating variations in responsiveness to level-of-service measures in travel mode choice modeling'. In: *Transportation Research Part A: Policy and Practice* 32.7, pp. 495–507.
Biljecki, F., Ledoux, H. and Oosterom, P. van (2013). 'Transportation mode-based segmentation and classification of movement trajectories'. In: *International Journal of Geographical Information Science* 27.2, pp. 385–407.
Bleisch, S., Duckham, M., Galton, A., Laube, P. and Lyon, J. (2014). 'Mining candidate causal relationships in movement patterns'. In: *International Journal of Geographical Information Science* 28.2, pp. 363–382.
Blondel, V. D., Decuyper, A. and Krings, G. (2015). 'A survey of results on mobile phone datasets analysis'. In: *Arxiv preprint*, arXiv:1502.03406v1.
Bohte, W. and Maat, K. (2009). 'Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands'. In: *Transportation Research Part C: Emerging Technologies* 17.3, pp. 285–297.
Bolbol, A. and Cheng, T. (2010). 'GPS Data Collection Setting For Pedestrian Activity Modelling'. In: *Proceedings of the GIS Research UK 18th Annual Conference GISRUK 2010*, pp. 337–344.
Bolbol, A., Cheng, T., Tsapakis, I. and Haworth, J. (2012). 'Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification'. In: *Computers, Environment and Urban Systems* 36.6, pp. 526–537.
Bricka, S. G., Sen, S., Paleti, R. and Bhat, C. R. (2012). 'An analysis of the factors influencing differences in survey-reported and GPS-recorded trips'. In: *Transportation Research Part C: Emerging Technologies* 21.1, pp. 67–88.
Bricka, S. and Bhat, C. (2006). 'Comparative analysis of global positioning system-based and travel survey-based data'. In: *Transportation Research Record: Journal of the Transportation Research Board* 1972, pp. 9–20.

Bricka, S., Zmud, J., Wolf, J. and Freedman, J. (2009). 'Household travel surveys with GPS: An experiment'. In: *Transportation Research Record: Journal of the Transportation Research Board* 2105, pp. 51–56.

Burkhard, O., Ahas, R., Saluveer, E. and Weibel, R. (2017). 'Extracting regular mobility patterns from sparse CDR data without a priori assumptions'. In: *Journal of Location Based Services* 11.2, pp. 78–97.

Cai, C., Gao, Y., Pan, L. and Zhu, J. (2015). 'Precise point positioning with quad-constellations: GPS, BeiDou, GLONASS and Galileo'. In: *Advances in Space Research* 56.1, pp. 133–143.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J. and Ratti, C. (2013). 'Understanding individual mobility patterns from urban sensing data: A mobile phone trace example'. In: *Transportation Research Part C: Emerging Technologies* 26, pp. 301–313.

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G. and Barabási, A.-L. (2007). 'Uncovering individual and collective human dynamics from mobile phone records'. In: *Journal of physics A: mathematical and theoretical* 41.22.

Chaloupka, Z. (2017). 'Technology and standardization gaps for high accuracy positioning in 5G'. In: *IEEE Communications Standards Magazine* 1.1, pp. 59–65.

Chandra, S. R. and Al-Deek, H. (2009). 'Predictions of Freeway Traffic Speeds and Volumes Using Vector Autoregressive Models'. In: *Journal of Intelligent Transportation Systems* 13.2, pp. 53–72.

Chang, C.-C. and Lin, C.-J. (2011). 'LIBSVM: A library for support vector machines'. In: *ACM Transactions on Intelligent Systems and Technology* 2.3, pp. 1–27.

Chen, C. H., Chang, Y. C., Chen, T. Y. and Wang, D. J. (2008). 'People counting system for getting in/out of a bus based on video processing'. In: *Proceedings - 8th International Conference on Intelligent Systems Design and Applications, ISDA 2008* 3, pp. 565–569.

Chen, C., Gong, H., Lawson, C. and Bialostozky, E. (2010). 'Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study'. In: *Transportation Research Part A: Policy and Practice* 44.10, pp. 830–840.

Chen, H., Rakha, H. A. and Sadek, S. (2011). 'Real-time freeway traffic state prediction: A particle filter approach'. In: *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on*, pp. 626–631.

Chen, L., Özsu, M. T. and Oria, V. (2005). 'Robust and fast similarity search for moving object trajectories'. In: *Proceedings of the 2005 ACM SIGMOD international conference on Management of data - SIGMOD '05*, p. 491.

Chen, Y.-y., Lv, Y., Li, Z. and Wang, F.-y. (2016). 'Long Short-Term Memory Model for Traffic Congestion Prediction with Online Open Data'. In: *19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 132–137.

Cho, K., Merrienboer, B. van, Bahdanau, D. and Bengio, Y. (2014). 'On the Properties of Neural Machine Translation: Encoder-Decoder Approaches'. In:

Chung, E. H. and Shalaby, A. (2005). 'A trip reconstruction tool for GPS-based personal travel surveys'. In: *Transportation Planning and Technology* 28.5, pp. 381–401.

Cooke, S. J., Nguyen, V. M., Kessel, S. T., Hussey, N. E., Young, N. and Ford, A. T. (2017). 'Troubling issues at the frontier of animal tracking for conservation and management'. In: *Conservation Biology* 31.5, pp. 1205–1207.

Csáji, B. C., Browet, A., Traag, V. a., Delvenne, J. C., Huens, E., Van Dooren, P., Smoreda, Z. and Blondel, V. D. (2013). 'Exploring the mobility of mobile phone users'. In: *Physica A: Statistical Mechanics and its Applications* 392.6, pp. 1459–1473.

Culnane, C., Rubinstein, B. I. and Teague, V. (2017). 'Privacy assessment of de-identified opal data: A report for transport for NSW'. In: *arXiv preprint arXiv:1704.08547*.

Dammann, A., Raulefs, R. and Zhang, S. (2015). 'On prospects of positioning in 5G'. In: *Communication Workshop (ICCW), 2015 IEEE International Conference on*. IEEE, pp. 1207–1213.

Das, R. D., Ronald, N. and Winter, S. (2014). 'Clustering based transfer detection with fuzzy activity recognition from smart-phone GPS trajectories'. In: *2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014*, pp. 3138–3143.

Das, R. D. and Winter, S. (2016). 'Detecting urban transport modes using a hybrid knowledge driven framework from GPS trajectory'. In: *ISPRS International Journal of Geo-Information* 5.11.

— (2018). 'A fuzzy logic based transport mode detection framework in urban environment'. In: *Journal of Intelligent Transportation Systems*, pp. 1–12.

De Montjoye, Y. A., Hidalgo, C. A., Verleysen, M. and Blondel, V. D. (2013). 'Unique in the Crowd: The privacy bounds of human mobility'. In: *Scientific Reports* 3.

Demšar, U., Buchin, K., Cagnacci, F., Safi, K., Speckmann, B., Van de Weghe, N., Weiskopf, D. and Weibel, R. (2015). 'Analysis and visualisation of movement: an interdisciplinary review'. In: *Movement Ecology* 3.1, p. 5.

Demšar, U., Slingsby, A. and Weibel, R. (2019). 'Introduction to the special section on Visual Movement Analytics'. In: *Information Visualization* 18.1, pp. 133–137.

Do, T. M. T. and Gatica-Perez, D. (2012). 'Contextual conditional models for smartphone-based human mobility prediction'. In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp)*, p. 163.

Dodge, S., Laube, P. and Weibel, R. (2012). 'Movement similarity assessment using symbolic representation of trajectories'. In: *International Journal of Geographical Information Science* 26.9, pp. 1563–1588.

Dodge, S., Weibel, R. and Lautenschütz, A.-K. (2008). 'Towards a taxonomy of movement patterns'. In: *Information Visualization* 7, pp. 240–252.

Doyle, J., Hung, P., Farrell, R. and McLoone, S. (2014). 'Population Mobility Dynamics Estimated from Mobile Telephony Data'. In: *Journal of Urban Technology* 21.2, pp. 109–132.

Dressler, F., Ripperger, S., Hierold, M., Nowak, T., Eibel, C., Cassens, B., Mayer, F., Meyer-Wegener, K. and Kolpin, A. (2016). 'From radio telemetry to ultra-low-power sensor networks: tracking bats in the wild'. In: *IEEE Communications Magazine* 54.1, pp. 129–135.

Eagle, N., De Montjoye, Y. A. and Bettencourt, L. M. A. (2009). 'Community computing: Comparisons between rural and urban societies using mobile phone data'. In: *Proceedings - 12th IEEE International Conference on Computational Science and Engineering, CSE 2009* 4, pp. 144–150.

Edelhoff, H., Signer, J. and Balkenhol, N. (2016). 'Path segmentation for beginners : an overview of current methods for detecting changes in animal movement patterns'. In: *Movement Ecology* 4.21.

El Raheb, K. and Ioannidis, Y. (2011). 'A labanotation based ontology for representing dance movement'. In: *International Gesture Workshop*, pp. 106–117.

Elhenawy, M. and Rakha, H. (2016). 'Stretch-Wide Traffic State Prediction Using Discriminatively Pre -Trained Deep Neural Networks'. In: *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, pp. 1065–1070.

Ellis, K., Godbole, S., Marshall, S., Lanckriet, G., Staudenmayer, J. and Kerr, J. (2014). 'Identifying Active Travel Behaviors in Challenging Environments Using GPS, Accelerometers, and Machine Learning Algorithms'. In: *Frontiers in Public Health* 2.April, pp. 1–8.

Ester, M., Kriegel, H. P., Sander, J. and Xu, X. (1996). 'A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'. In: *Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231.

European Emergency Number Association (2011). *Caller Location in Support of Emergency Services*. Tech. rep.

Farrahi, K. and Gatica-Perez, D. (2010). 'Probabilistic mining of socio-geographic routines from mobile phone data'. In: *IEEE Journal on Selected Topics in Signal Processing* 4.4, pp. 746–755.

Feng, T. and Timmermans, H. J. (2016). 'Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data'. In: *Transportation Planning and Technology* 39.2, pp. 180–194.

Ferrara, A., Sacone, S. and Siri, S. (2018). *Freeway Traffic Modelling and Control*. Springer.

Ferster, C. J., Fischer, J., Manaugh, K., Nelson, T. and Winters, M. (2018). 'Using OpenStreetMap to Inventory Bicycle Infrastructure: A Comparison with Open Data from Cities'. In: *Transportation Research Board 97th Annual Meeting*.

Fillekes, M. P., Röcke, C., Katana, M. and Weibel, R. (2019). 'Self-reported versus GPS-derived indicators of daily mobility in a sample of healthy older adults'. In: *Social Science & Medicine* 220, pp. 193–202.

Forrest, T. and Pearson, D. (2005). 'Comparison of trip determination methods in household travel surveys enhanced by a Global Positioning System'. In: *Journal of the Transportation Research Board* 1917, pp. 63–71.

Fu, R., Zhang, Z. and Li, L. (2016). 'Using LSTM and GRU neural network methods for traffic flow prediction'. In: *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, pp. 324–328.

Furletti, B., Gabrielli, L., Rinzivillo, S. and Renso, C. (2012). 'Identifying users profiles from mobile calls habits'. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing . ACM*. Pp. 17–24.

Furletti, B., Cnr, K. I., Cintia, P., Cnr, K.-. I. and Spinsanti, L. (2013). 'Inferring human activities from GPS tracks'. In: *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, p. 5.

Geddes, A. and Scholten, P. (2016). *The politics of migration & immigration in Europe*. Los Angeles: Sage.

Gerland, H. E. and Sutter, K. (1999). 'Automatic passenger counting (apc): Infrared motion analyzer for accurate counts in stations and rail, light-rail and bus operations'. In: *APTA Bus Conference*.

Givoni, M. and Perl, A. (2017). 'Rethinking Transport Infrastructure Planning to Extend Its Value over Time'. In: *Journal of Planning Education and Research*, p. 0739456X1774119.

Gong, H., Chen, C., Bialostozky, E. and Lawson, C. T. (2012). 'A GPS/GIS method for travel mode detection in New York City'. In: *Computers, Environment and Urban Systems* 36.2, pp. 131–139.

Gong, L., Morikawa, T., Yamamoto, T. and Sato, H. (2014). 'Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies'. In: *Procedia - Social and Behavioral Sciences* 138, pp. 557–565.

Gonzalez, M. C., Hidalgo, C. A. and Barabási, A.-L. (2008). 'Understanding individual human mobility patterns'. In: *Nature* 453, pp. 779–782.

Gonzalez, P. A., Weinstein, J. S., Barbeau, S. J., Labrador, M. A., Winters, P. L., Georggi, N. L. and Perez, R. A. (2008). 'Automating Mode Detection Using Neural Networks and Assisted GPS Data Collected Using GPS-Enabled Mobile Phones'. In: *15th World Congress on Intelligent Transport Systems*, 12p.

Goodchild, M. F. (2007). 'Citizens as sensors: The world of volunteered geography'. In: *GeoJournal* 69.4, pp. 211–221.

Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. MIT Press.

Gössling, S. (2013). 'Urban transport transitions: Copenhagen, city of cyclists'. In: *Journal of Transport Geography* 33, pp. 196–206.

Gurarie, E., Bracis, C., Delgado, M., Meckley, T. D., Kojola, I. and Wagner, C. M. (2016). 'What is the animal doing ? Tools for exploring behavioural structure in animal movements'. In: *Journal of Animal Ecology* 85.1, pp. 69–84.

Haegerstrand, T. (1970). 'What about people in Regional Science?' In: *Papers in Regional Science* 24.6.

Hahsler, M., Grün, B. and Hornik, K. (2005). 'Association Rules and Frequent Item Sets'. In: *Journal of Statistical Software* 14.15.

Hamdy, Y. R. and Mawjoud, S. A. (2012). 'Performance assessment of U-TDOA and A-GPS positioning methods'. In: *2012 International Conference on Future Communication Networks, ICFCN 2012*, pp. 99–104.

Hastie, T., Tibshirani, R. and Friedman, J. (2017). *The Elements of Statistical Learning*.

He, K., Zhang, X., Ren, S. and Sun, J. (2016). 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.

Helbing, D., Buzna, L., Johansson, A. and Werner, T. (2005). 'Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions'. In: *Transportation science* 39.1.

Hemminki, S., Nurmi, P. and Tarkoma, S. (2013). 'Accelerometer-based transportation mode detection on smartphones'. In: *Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems - SenSys '13*, pp. 1–14.

Hochreiter, S. and Schmidhuber, J. J. (1997). 'Long short-term memory'. In: *Neural Computation* 9.8, pp. 1–32.

Horn, C., Klampfl, S., Cik, M. and Reiter, T. (2014). 'Detecting outliers in cell phone data: correcting trajectories to improve traffic modeling'. In: *Transportation Research Record: Journal of the Transportation Research Board* 2405, pp. 49–56.

Hoshiar, A. K., Le, T.-A., Amin, F. U., Kim, M. O. and Yoon, J. (2017). 'Studies of aggregated nanoparticles steering during magnetic-guided drug delivery in the blood vessels'. In: *Journal of Magnetism and Magnetic Materials* 427, pp. 181–187.

Houston, D., Luong, T. T. and Boarnet, M. G. (2014). 'Tracking daily travel; Assessing discrepancies between GPS-derived and self-reported travel patterns'. In: *Transportation Research Part C: Emerging Technologies* 48, pp. 97–108.

Huang, Z., Xu, W. and Yu, K. (2015). 'Bidirectional LSTM-CRF models for sequence tagging'. In: *arXiv preprint arXiv:1508.01991*.

Huss, A., Beekhuizen, J., Kromhout, H. and Vermeulen, R. (2014). 'Using GPS-derived speed patterns for recognition of transport modes in adults'. In: *International Journal of Health Geographics* 13.1, p. 40.

Isaacman, S., Becker, R., Caceres, R., Kobourov, S., Martonosi, M., Rowland, J. and Varshavsky, A. (2011). 'Identifying important places in people's lives from cellular network data'. In: *Lecture Notes in Computer Science* 6696.June, pp. 133–151.

Jahangiri, A. and Rakha, H. A. (2015). 'Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data'. In: *IEEE Transactions on Intelligent Transportation Systems* 16.5, pp. 2406–2417.

Janecek, A., Valerio, D., Hummel, K. A., Ricciato, F. and Hlavacs, H. (2015). 'The Cellular Network as a Sensor: From Mobile Phone Data to Real-Time Road Traffic Monitoring'. In: *IEEE Transactions on Intelligent Transportation Systems* 16.5, pp. 2551–2572.

Janzen, M., Vanhoof, M., Axhausen, K. W. and Smoreda, Z. (2016). 'Estimating Long-Distance Travel Demand with Mobile Phone Billing Data'. In: *16th Swiss Transport Research Conference (STRC 2016)*. Monte Verità.

Jiang, S., Fiore, G., Yang, Y. and Ferreira, J. J. (2013a). 'A review of urban computing for mobile phone traces: current methods, challenges and opportunities'. In: *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*.

Jiang, S., Ferreira, J. and Gonzalez, M. C. (2017). 'Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore'. In: *IEEE Transactions on Big Data* 3.2, pp. 208–219.

Jiang, S., Ferreira, J. and Gonzalez, M. C. (2012). 'Clustering daily patterns of human activities in the city'. In: *Data Mining and Knowledge Discovery* 25.3, pp. 478–510.

Jiang, Z.-Q., Xie, W.-J., Li, M.-X., Podobnik, B., Zhou, W.-X. and Stanley, H. E. (2013b). 'Calling patterns in human communication dynamics'. In: *Proceedings of the National Academy of Sciences* 110.5, pp. 1600–1605.

Jony, R. I., Habib, A., Mohammed, N. and Rony, R. I. (2015). 'Big data use case domains for telecom operators'. In: *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pp. 850–855.

Kamarianakis, Y. and Prastacos, P. (2003). 'Forecasting Traffic Flow Conditions in an Urban Network: Comparison of Multivariate and Univariate Approaches'. In: *Transportation Research Record* 1857, pp. 74–84.

Kays, R. et al. (2011). 'Tracking animal location and activity with an automated radio telemetry system in a tropical rainforest'. In: *The Computer Journal* 54.12, pp. 1931–1948.

Kerr, J., Duncan, S. and Schipperjin, J. (2011). 'Using global positioning systems in health research: A practical approach to data collection and processing'. In: *American Journal of Preventive Medicine* 41.5, pp. 532–540.

Kiukkonen, N., Blom, J., Dousse, O., Gatica-Perez, D. and Laurila, J. (2010). 'Towards rich mobile phone datasets: Lausanne data collection campaign'. In: *Proc. ICPS, Berlin*.

Knoblauch, R., Pietrucha, M. and Nitzburg, M. (1996). 'Field studies of pedestrian walking speed and start-up time'. In: *Transportation Research Record: Journal of the Transportation Research Board* 1538, pp. 27–38.

Knospe, W., Santen, L., Schadschneider, A. and Schreckenberg, M. (2000). 'Towards a realistic microscopic description of highway traffic'. In: *Journal of Physics A: Mathematical and general* 33.48.

Koivisto, M., Hakkarainen, A., Costa, M., Talvitie, J., Heiska, K., Leppanen, K. and Valkama, M. (2017). 'Continuous high-accuracy radio positioning of cars in ultra-dense 5G networks'. In: *2017 13th International Wireless Communications and Mobile Computing Conference, IWCMC 2017*, pp. 115–120.

Kok, A. L., Hans, E. W. and Schutten, J. M. (2012). 'Vehicle routing under time-dependent travel times: The impact of congestion avoidance'. In: *Computers and Operations Research* 39.5, pp. 910–918.

Krumm, J., Rouhana, D. and Chang, M. W. (2015). 'Placer++: Semantic place labels beyond the visit'. In: *2015 IEEE International Conference on Pervasive Computing and Communications, PerCom 2015*, pp. 11–19.

Kung, K. S., Greco, K., Sobolevsky, S. and Ratti, C. (2014). 'Exploring universal patterns in human home-work commuting from mobile phone data'. In: *PLoS ONE* 9.6.

Lara, O. D. and Labrador, M. A. (2013). 'A Survey on Human Activity Recognition using Wearable Sensors'. In: *IEEE Communications Surveys & Tutorials* 15.3, pp. 1192–1209.

Laube, P. (2014). *Computational movement analysis*. Springer.

Laube, P. and Purves, R. S. (2011). 'How fast is a cow? Cross-Scale Analysis of Movement Data'. In: *Transactions in GIS* 15.3, pp. 401–418.

Lenntorp, B. (1999). 'Time-geography - At the end of its beginning'. In: *GeoJournal* 48.3, pp. 155–158.

Leuzzi, F., Del Signore, E. and Ferranti, R. (2017). 'Towards a Pervasive and Predictive Traffic Police'. In: *Italian Conference for the Traffic Police*. Springer, pp. 19–35.

Liao, L., Fox, D. and Kautz, H. (2007). 'Learning and Inferring Transportation Routines'. In: *Artificial Intelligence* 171.5-6, pp. 311–331.

Lighthill, M. J. and Whitham, G. B. (1955). 'On Kinematic Waves. II. A Theory of Traffic Flow on Long Crowded Roads'. In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 229.1178, pp. 317–345.

Lin, M., Hsu, W.-J. and Lee, Z. Q. (2013). 'Detecting modes of transport from unlabeled positioning sensor data'. In: *Journal of Location Based Services* 7.4, pp. 272–290.

Lin, Z., Yin, M., Feygin, S., Sheehan, M., Paiment, J.-F. and Pozdnoukhov, A. (2017). 'Deep Generative Models of Urban Mobility'. In: *IEEE Transactions on Intelligent Transportation Systems*.

Linnap, M. and Rice, A. (2014). 'Managed Participatory Sensing with YouSense'. In: *Journal of Urban Technology* 21.2, pp. 9–26.

Liu, L., Biderman, A. and Ratti, C. (2009). 'Urban Mobility Landscape : Real Time Monitoring of Urban Mobility Patterns'. In: *Proceedings of the 11th International Conference on Computers in Urban Planning and Urban Management*, pp. 1–16.

Liu, Y., Zheng, H., Feng, X. and Chen, Z. (2017). 'Short-term traffic flow prediction with Conv-LSTM'. In: *2017 9th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6.

Louail, T., Lenormand, M., Cantu Ros, O. G., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J. J. and Barthelemy, M. (2014). 'From mobile phone data to the spatial structure of cities.' In: *Scientific reports* 4, p. 5276.

Lu, X., Wetter, E., Bharti, N., Tatem, A. J. and Bengtsson, L. (2013). 'Approaching the limit of predictability in human mobility.' In: *Scientific reports* 3, p. 2923.

Lu, X. et al. (2016). 'Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh'. In: *Global Environmental Change* 38, pp. 1–7.

Lv, M., Chen, L. and Chen, G. (2012). 'Discovering personally semantic places from GPS trajectories'. In: *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, p. 1552.

Lv, Y., Duan, Y., Kang, W., Li, Z. and Wang, F. Y. (2014). 'Traffic Flow Prediction With Big Data: A Deep Learning Approach'. In: *IEEE Transactions on Intelligent Transportation Systems* 16.2, pp. 865–873.

Ma, X., Tao, Z., Wang, Y., Yu, H. and Wang, Y. (2015). 'Long short-term memory neural network for traffic speed prediction using remote microwave sensor data'. In: *Transportation Research Part C: Emerging Technologies* 54, pp. 187–197.

Mäenpää, H., Lobov, A. and Martinez Lastra, J. L. (2017). 'Travel mode estimation for multi-modal journey planner'. In: *Transportation Research Part C: Emerging Technologies* 82, pp. 273–289.

Marcos, D., Volpi, M., Kellenberger, B. and Tuia, D. (2018). 'Land cover mapping at very high resolution with rotation equivariant CNNs: towards small yet accurate models'. In: *CoRR* abs/1803.0.

Miller, H. J., Dodge, S., Miller, J. and Bohrer, G. (2019). 'Towards an integrated science of movement: converging research on animal movement ecology and human mobility science'. In: *International Journal of Geographical Information Science*.

Modsching, M., Kramer, R. and Hagen, K. (2006). 'Field trial on GPS Accuracy in a medium size city: The influence of built-up'. In: *3rd workshop on positioning, navigation and communication*, pp. 209–218.

Moiseeva, A. and Timmermans, H. (2010). 'Imputing relevant information from multi-day GPS tracers for retail planning and management using data fusion and context-sensitive learning'. In: *Journal of Retailing and Consumer Services* 17.3, pp. 189–199.

Müller, P., Del Peral-Rosado, J., Robert, P. and Seco-Granados, G. (2016). 'Statistical Trilateration With Skew-t Distributed Errors in LTE Networks'. In: *IEEE Transactions on Wireless Communications* 15.10, pp. 7114–7127.

Nantes, A., Ngoduy, D., Bhaskar, A., Miska, M. and Chung, E. (2016). 'Real-time traffic state estimation in urban corridors from heterogeneous data'. In: *Transportation Research Part C: Emerging Technologies* 66, pp. 99–118.

Nathan, R., Spiegel, O., Fortmann-Roe, S., Harel, R., Wikelski, M. and Getz, W. M. (2012). 'Using tri-axial acceleration data to identify behavioral modes of free-ranging animals: general concepts and tools illustrated for griffon vultures'. In: *Journal of Experimental Biology* 215.6, pp. 986–996.

Neutens, T. (2015). 'Accessibility, equity and health care: Review and research directions for transport geographers'. In: *Journal of Transport Geography* 43, pp. 14–27.

Ni, L., Wang, X. C. and Chen, X. M. (2018). 'A spatial econometric model for travel flow analysis and real-world applications with massive mobile phone data'. In: *Transportation Research Part C: Emerging Technologies* 86, pp. 510–526.

Nilbe, K., Ahas, R. and Slim, S. (2014). 'Evaluating the Travel Distances of Events Visitors and Regular Visitors Using Mobile Positioning Data: The Case of Estonia'. In: *Journal of Urban Technology* 21.2, pp. 91–107.

Nitsche, P., Widhalm, P., Breuss, S., Brändle, N. and Maurer, P. (2014). 'Supporting large-scale travel surveys with smartphones – A practical approach'. In: *Transportation Research Part C* 43, pp. 212–221.

Nitsche, P., Widhalm, P., Breuss, S. and Maurer, P. (2012). 'A strategy on how to utilize smartphones for automatically reconstructing trips in travel surveys'. In: *Procedia - Social and Behavioral Sciences* 48, pp. 1033–1046.

Nyhan, M. et al. (2016). 'Predicting vehicular emissions in high spatial resolution using pervasively measured transportation data and microscopic emissions model'. In: *Atmospheric Environment* 140, pp. 352–363.

Okutani, I. and Stephanedes, Y. J. (1984). 'Dynamic prediction of traffic volume through Kalman filtering theory'. In: *Transportation Research Part B: Methodological* 18.1, pp. 1–11.

Papageorgiou, M., Diakaki, C., Dinopoulou, V., Kotsialos, A. and Wang, Y. (2003). 'Review of Road Traffic Control Strategies'. In: *Proc. IEEE* 91.12, pp. 2043–2067.

Pappalardo, L., Simini, F., Rinzivillo, S., Pedreschi, D. and Giannotti, F. (2013). 'Comparing general mobility and mobility by car'. In: *11th Brazilian Congress on Computational Intelligence (BRICS-CCI CBIC)*, pp. 665–668.

Parent, C. et al. (2013). 'Semantic trajectories modeling and analysis'. In: *ACM Computing Surveys* 45.4, 42:1–42:32.

Park, B. D., Rilett, L. R. and Han, G. (1999). 'Spectral Basis Neural Networks For Real-Time Travel Time Forecasting'. In: *Journal of Transportation Engineering* 3.December, pp. 515–523.

Patier, D. and Routhier, J.-L. (2009). 'How to improve the capture of urban goods movement data?' In: *Transport Survey Methods: Keeping up with a Changing World*, pp. 251–287.

Pedregosa, F. et al. (2011). 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peral-Rosado, J. A. del, López-Salcedo, J. A., Zanier, F. and Crisci, M. (2012). 'Achievable localization accuracy of the positioning reference signal of 3GPP LTE'. In: *Localization and GNSS (ICL-GNSS), 2012 International Conference on*. IEEE, pp. 1–6.

Pereira, F., Carrion, C., Zhao, F., Cottrill, C. D., Zegras, C. and Ben-Akiva, M. E. (2013). 'The Future Mobility Survey: overview and preliminary evaluation'. In: *Proceedings of the Eastern Asia Society for Transporation Studies* 9.

Plank, B., Søgaard, A. and Goldberg, Y. (2016). 'Multilingual Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Models and Auxiliary Loss'. In: *CoRR* abs/1604.0.

Polson, N. G. and Sokolov, V. O. (2017). 'Deep learning for short-term traffic flow prediction'. In: *Transportation Research Part C: Emerging Technologies* 79, pp. 1–17.

Poppe, R. (2010). 'A survey on vision-based human action recognition'. In: *Image and Vision Computing* 28.6, pp. 976–990.

Prelipcean, A. C., Gidofalvi, G. and Susilo, Y. O. (2016). 'Measures of transport mode segmentation of trajectories'. In: *International Journal of Geographical Information Science* 8816.March, pp. 1–22.

Prelipcean, A. C., Gidófalvi, G. and Susilo, Y. O. (2017). 'Transportation mode detection – an in-depth review of applicability and reliability applicability and reliability'. In: *Transport Reviews* 37.4, pp. 442–464.

Ranacher, P., Brunauer, R., Trutschnig, W., Van der Spek, S. and Reich, S. (2016). 'Why GPS makes distances bigger than they are'. In: *International Journal of Geographical Information Science* 30.2, pp. 316–333.

Ranjan, G., Zang, H., Zhang, Z.-L. and Bolot, J. (2012). 'Are call detail records biased for sampling human mobility?' In: *ACM SIGMOBILE Mobile Computing and Communications Review* 16.3, p. 33.

Reades, J., Calabrese, F. and Ratti, C. (2009). 'Eigenplaces: Analysing cities using the space - Time structure of the mobile phone network'. In: *Environment and Planning B: Planning and Design* 36.5, pp. 824–836.

Reddy, S., Burke, J., Estrin, D., Hansen, M. and Srivastava, M. (2008). 'Determining transportation mode on mobile phones'. In: *Proceedings - International Symposium on Wearable Computers, ISWC*, pp. 25–28.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M. and Srivastava, M. (2010). 'Using mobile phones to determine transportation modes'. In: *ACM Transactions on Sensor Networks* 6.2, pp. 1–27.

Renso, C., Baglioni, M., Macedo, J. A. F. de, Trasarti, R. and Wachowicz, M. (2013). 'How you move reveals who you are: understanding human behavior by analyzing trajectory data'. In: *Knowledge and Information Systems* 37.2, pp. 331–362.

Rinzivillo, S., Gabrielli, L., Nanni, M., Pappalardo, L., Pedreschi, D. and Giannotti, F. (2014). 'The Purpose of Motion : Learning Activities from Individual Mobility Networks'. In: *International Conference on Data Science and Advanced Analytics (DSAA14)*.

Rinzivillo, S., Mainardi, S., Pezzoni, F., Coscia, M., Pedreschi, D. and Giannotti, F. (2012). 'Discovering the Geographical Borders of Human Mobility'. In: *KI Künstliche Intelligenz* 26.3, pp. 253–260.

Rumelhart, D. E., Hinton, G. E. and Williams, R. J. (1985). *Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Institute for Cognitive Science.

Rutherford, S. G., McCormack, E. and Wilkinson, M. (1996). 'Travel Impacts of Urban Form: Implications From an Analysis of Two Seattle Area Travel Diaries. TMIP Conference on Urban Design, Telecommuting, and Travel Behaviour, Washington, D.C.' In: *Urban Design, Telecommuting and Travel Forecasting Conference*, pp. 95–166.

Saad, S., De Beul, D., Mahmoudi, S. and Manneback, P. (2012). 'An Ontology for video human movement representation based on Benesh notation'. In: *Proceedings of 2012 International Conference on Multimedia Computing and Systems, ICMCS 2012*, pp. 77–82.

Saeb, S., Lattie, E. G., Schueller, S. M., Kording, K. P. and Mohr, D. C. (2016). 'The relationship between mobile phone location sensor data and depressive symptom severity'. In: *PeerJ* 4, e2537.

Sauerländer-Biebl, A., Brockfeld, E., Suske, D. and Melde, E. (2017). 'Evaluation of a transport mode detection using fuzzy rules'. In: *Transportation Research Procedia* 25, pp. 591–602.

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z. and González, M. C. (2013). 'Unravelling daily human mobility motifs.' In: *Journal of the Royal Society Interface* 10.84.

Schuessler, N. and Axhausen, K. W. (2009). 'Processing Raw Data from Global Positioning Systems Without Additional Information'. In: *Transportation Research Record: Journal of the Transportation Research Board* 2105, pp. 28–36.

Schulz, D., Bothe, S. and Korner, C. (2012). 'Human Mobility from GSM Data - A Valid Alternative to GPS ?' In: *Mobile data challenge 2012 workshop, June*, pp. 18–19.

Sejnowski, T. J. and Rosenberg, C. R. (1987). 'Parallel networks that learn to pronounce English text'. In: *Complex systems* 1.1, pp. 145–168.

Semanjski, I., Gautama, S., Ahas, R. and Witlox, F. (2017). 'Spatial context mining approach for transport mode recognition from mobile sensed big data'. In: *Computers, Environment and Urban Systems* 66, pp. 38–52.

Shafique, M. A. and Hato, E. (2015). 'Use of acceleration data for transportation mode prediction'. In: *Transportation* 42.1, pp. 163–188.
Shah, R. C., Wan, C.-y., Lu, H. and Nachman, L. (2014). 'Classifying the mode of transportation on mobile phones using GIS information'. In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct*, pp. 225–229.
Shahsavari, B. and Abbeel, P. (2015). *Short-Term Traffic Forecasting: Modeling and Learning Spatio-Temporal Relations in Transportation Networks Using Graph Neural Networks*. Tech. rep. University of California, Berkeley.
Shao, H. and Soong, B.-H. (2016). 'Traffic flow prediction with Long Short-Term Memory Networks (LSTMs)'. In: *Region 10 Conference (TENCON), 2016 IEEE*, pp. 2986–2989.
Shen, L. and Stopher, P. R. (2014). 'Review of GPS Travel Survey and GPS Data-Processing Methods'. In: *Transport Reviews* 34.3, pp. 316–334.
Sheu, J. B., Lan, L. W. and Huang, Y. S. (2009). 'Short-term prediction of traffic dynamics with real-time recurrent learning algorithms'. In: *Transportmetrica* 5.1, pp. 59–83.
Shumway, R. H. and Stoffer, D. S. (2010). *Time Series Analysis and Its Applications: With R Examples*. Springer T. Springer.
— (2017). *Time Series Analysis and Its Applications*. Springer.
Silm, S. and Ahas, R. (2014). 'The temporal variation of ethnic segregation in a city: Evidence from a mobile phone use dataset.' In: *Social science research* 47, pp. 30–43.
Smith, B. L. and Demetsky, M. J. (1994). 'Short-term traffic flow prediction models-a comparison of neural network and nonparametric regression approaches'. In: *IEEE International Conference on Systems, Man, and Cybernetics*, pp. 1706–1709.
Smouse, P. E., Focardi, S., Moorcroft, P. R., Kie, J. G., Forester, J. D. and Morales, J. M. (2010). 'Stochastic modelling of animal movement'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 365.1550, pp. 2201–2211.
Soleymani, A., Cachat, J., Robinson, K., Dodge, S., Kalueff, A. and Weibel, R. (2014). 'Integrating cross-scale analysis in the spatial and temporal domains for classification of behavioral movement'. In: *Journal of Spatial Information Science* 8.
Soleymani, A., Pennekamp, F., Dodge, S. and Weibel, R. (2017). 'Characterizing change points and continuous transitions in movement behaviours using wavelet decomposition'. In: *Methods in Ecology and Evolution* 8.9, pp. 1113–1123.
Song, C., Qu, Z., Blumm, N. and Barabási, A.-L. (2010). 'Limits of predictability in human mobility.' In: *Science (New York, N.Y.)* 327.5968, pp. 1018–1021.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). 'Dropout: A Simple Way to Prevent Neural Networks from Overfitting'. In: *Journal of Machine Learning Research* 15, pp. 1929–1958.
Stathopoulos, A. and Karlaftis, M. G. (2003). 'A multivariate state space approach for urban traffic flow modeling and prediction'. In: *Transportation Research Part C: Emerging Technologies* 11.2, pp. 121–135.

Steenbruggen, J., Tranos, E. and Nijkamp, P. (2015). 'Data from mobile phone operators: A tool for smarter cities?' In: *Telecommunications Policy* 39.3-4, pp. 335–346.

Stenneth, L., Wolfson, O., Yu, P. S. and Xu, B. (2011). 'Transportation mode detection using mobile phones and GIS information'. In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 54.

Stopher, P. R. and Greaves, S. P. (2007). 'Household travel surveys: Where are we going?' In: *Transportation Research Part A: Policy and Practice* 41.5, pp. 367–381.

Stopher, P., FitzGerald, C. and Zhang, J. (2008). 'Search for a global positioning system device to measure person travel'. In: *Transportation Research Part C: Emerging Technologies* 16.3, pp. 350–369.

Sun, C., Shrivastava, A., Singh, S. and Gupta, A. (2017). 'Revisiting Unreasonable Effectiveness of Data in Deep Learning Era'. In: *Proceedings of the IEEE International Conference on Computer Vision* 2017-Octob, pp. 843–852.

Sun, S., Zhang, C. and Yu, G. (2006). 'A Bayesian Network Approach to Traffic Flow Forecasting'. In: *IEEE Transactions on Intelligent Transportation Systems* 7.1, pp. 124–132.

Sutskever, I., Vinyals, O. and Le, Q. V. (2014). 'Sequence to Sequence Learning with Neural Networks'. In: *NIPS*, pp. 3104–3112.

Szeto, W. Y., Ghosh, B., Basu, B. and Mahony, M. O. (2009). 'Cell Transmission Model and SARIMA Model'. In: *Journal of Transportational Engineering* 135.September, pp. 658–667.

Tanahashi, Y., Rowland, J., North, S. and Ma, K.-L. (2012). 'Inferring human mobility patterns from anonymized mobile communication usage'. In: *Proceedings of the 10th International Conference on Advances in Mobile Computing & Multimedia - MoMM '12*, p. 151.

Tenkanen, H., Saarsalmi, P., Järv, O., Salonen, M. and Toivonen, T. (2016). 'Health research needs more comprehensive accessibility measures: Integrating time and transport modes from open data'. In: *International Journal of Health Geographics* 15.1, pp. 1–12.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D. and Dorsett, K. (2015). 'Changes in data sharing and data reuse practices and perceptions among scientists worldwide'. In: *PloS one* 10.8, e0134826.

Thomson, J., Börger, L., Christianen, M., Esteban, N., Laloe, J.-O. and Hays, G. (2017). 'Implications of location accuracy and data volume for home range estimation and fine-scale movement analysis: comparing Argos and Fastloc-GPS tracking data'. In: *Marine Biology* 164.10.

Tian, Y. and Pan, L. (2015). 'Predicting Short-Term Traffic Flow by Long Short-Term Memory Recurrent Neural Network'. In: *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, pp. 153–158.

Toledo, S., Orchan, Y., Shohami, D., Charter, M. and Nathan, R. (2018). 'Physical-Layer Protocols for Lightweight Wildlife Tags with Internet-of-Things Transceivers'. In: *19th IEEE International Symp. on the World of Wireless, Mobile, and Multimedia Networks*.

Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A. and González, M. C. (2015). 'The path most traveled: Travel demand estimation using big data resources'. In: *Transportation Research Part C: Emerging Technologies* 58, pp. 192–177.

Trasarti, R., Olteanu-Raimond, A.-M., Nanni, M., Couronné, T., Furletti, B., Giannotti, F., Smoreda, Z. and Ziemlicki, C. (2015). 'Discovering urban and country dynamics from mobile phone data with spatial correlation patterns'. In: *Telecommunications Policy* 39.3-4, pp. 347–362.

Trevisani, E. and Vitaletti, A. (2004). 'Cell-ID location technique, limits and benefits: An experimental study'. In: *Proceedings - IEEE Workshop on Mobile Computing Systems and Applications, WMCSA* Wmcsa, pp. 51–60.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B. and Zafeiriou, S. (2016). 'Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network'. In: *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 5200–5204.

Tselentis, D. I., Yannis, G. and Vlahogianni, E. I. (2016). 'Innovative Insurance Schemes: Pay as/how You Drive'. In: *Transportation Research Procedia* 14, pp. 362–371.

Umair, M., Kim, W. S., Choi, B. C. and Jung, S. Y. (2014). 'Discovering personal places from location traces'. In: *16th International Conference on Advanced Communication Technology*, pp. 709–713.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017). 'Attention Is All You Need'. In: *CoRR* abs/1706.0.

Vij, A. and Shankari, K. (2015). 'When is big data big enough? Implications of using GPS-based surveys for travel demand analysis'. In: *Transportation Research Part C: Emerging Technologies* 56, pp. 446–462.

Vlahogianni, E. I., Karlaftis, M. G. and Golias, J. C. (2014). 'Short-term traffic forecasting: Where we are and where we're going'. In: *Transportation Research Part C: Emerging Technologies* 43, pp. 3–19.

Vlahogianni, E. I., Golias, J. C. and Karlaftis, M. G. (2004). 'Short-term traffic forecasting: Overview of objectives and methods'. In: *Transport Reviews* 24.5, pp. 533–557.

Vlahogianni, E. I., Karlaftis, M. G. and Golias, J. C. (2005). 'Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach'. In: *Transportation Research Part C: Emerging Technologies* 13.3, pp. 211–234.

Wakefield, E. D., Phillips, R. A. and Matthiopoulos, J. (2009). 'Quantifying habitat use and preferences of pelagic seabirds using individual movement data: a review'. In: *Marine Ecology Progress Series* 391, pp. 165–182.

Wang, H., Calabrese, F., Di Lorenzo, G. and Ratti, C. (2010). 'Transportation mode inference from anonymized and aggregated mobile phone call detail records'. In: *IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC*, pp. 318–323.

Wang, Y. and Nihan, N. L. (2003). 'Can Single-Loop Detectors Do the Work of Dual-Loop Detectors?' In: *Journal of Transportation Engineering* 129.2, pp. 169–176.

Weide, R. (2005). *The Carnegie mellon pronouncing dictionary [cmudict. 0.6]*. Tech. rep.

Whittaker, J., Garside, S. and Lindveld, K. (1997). 'Tracking and predicting a network traffic process'. In: *International Journal of Forecasting* 13.1, pp. 51–61.

Widhalm, P., Nitsche, P., Br, N., Widhalm, P., Nitsche, P. and Braendle, N. (2012). 'Transport Mode Detection with Realistic Smartphone Sensor Data'. In: Icpr, pp. 573–576.

Widhalm, P., Yang, Y., Ulm, M., Athavale, S. and González, M. C. (2015). 'Discovering urban activity patterns in cell phone data'. In: *Transportation* 42.4, pp. 597–623.

Williams, B. M. and Hoel, L. A. (2003). 'Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results'. In: *Journal of Transportation Engineering* 129.6, pp. 664–672.

Wolf, J., Guensler, R. and Bachman, W. (2001). 'Elimination of the Travel Diary: Experiment to Derive Trip Purpose from Global Positioning System Travel Data'. In: *Transportation Research Record: Journal of the Transportation Research Board* 1768.August 2016, pp. 125–134.

Wolf, J., Guensler, R., Washington, S. and Frank, L. (1999). 'Ues of electronic travel diaries and vehicle instrumentation packages in the year 2000'. In: *TRB Transportation Research Circular*, pp. 413–429.

Wu, Y. and Tan, H. (2016). 'Short-term traffic flow forecasting with spatio-temporal correlation in a hybrid deep learning framework'. In: *CoRR*, arXiv:1612.01022.

Xiao, G., Juan, Z. and Zhang, C. (2015). 'Travel mode detection based on GPS track data and Bayesian networks'. In: *Computers, Environment and Urban Systems* 54, pp. 14–22.

Xie, Y. and Zhang, Y. (2006). 'A Wavelet Network Model for Short-Term Traffic Volume Forecasting'. In: *Journal of Intelligent Transportation Systems: Technology, Planning, and Operations* 10.3, pp. 141–150.

Ye, Y., Zheng, Y., Chen, Y., Feng, J. and Xie, X. (2009). 'Mining individual life pattern based on location history'. In: *Proceedings - IEEE International Conference on Mobile Data Management*, pp. 1–10.

Yin, H., Wong, S. C., Xu, J. and Wong, C. K. (2002). 'Urban traffic flow prediction using a fuzzy-neural approach'. In: *Transportation Research Part C: Emerging Technologies* 10.2, pp. 85–98.

Yu, B., Yin, H. and Zhu, Z. (2017). 'Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting'. In: *CoRR*.

Yu, H., Wu, Z., Wang, S., Wang, Y. and Ma, X. (2017). 'Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks'. In: *Sensors (Switzerland)* 17.7, pp. 1–16.

Zandbergen, P. A. (2009). 'Accuracy of iPhone locations: A comparison of assisted GPS, WiFi and cellular positioning'. In: *Transactions in GIS* 13.SUPPL. 1, pp. 5–25.

Zaremba, W., Sutskever, I. and Vinyals, O. (2014). 'Recurrent Neural Network Regularization'. In: *arXiv preprint arXiv:1409.2329*.

Zeisl, B., Sattler, T. and Pollefeys, M. (2015). 'Camera Pose Voting for Large-Scale Image-Based Localization'. In: *The IEEE International Conference on Computer Vision (ICCV)*.

Zhang, L., Dalyot, S., Eggert, D. and Sester, M. (2012). 'Multi-Stage Approach To Travel-Mode Segmentation and Classification of Gps Traces'. In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XXXVIII-4/.October, pp. 87–93.

Zhao, Z., Chen, W., Wu, X., Chen, P. C. Y. and Liu, J. (2017). 'LSTM network: a deep learning approach for short-term traffic forecast'. In: *IET Intelligent Transport Systems* 11.2, pp. 68–75.

Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J. and Yin, L. (2016). 'Understanding the bias of call detail records in human mobility research'. In: *International Journal of Geographical Information Science* 8816.January, pp. 1–25.

Zheng, Y., Liu, L., Wang, L. and Xie, X. (2008). 'Learning transportation mode from raw gps data for geographic applications on the web'. In: *Proceeding of the 17th international conference on World Wide Web - WWW '08*, p. 247.

Zheng, Y., Xie, X. and Ma, W.-Y. (2010). 'Geolife: A collaborative social networking service among user, location and trajectory'. In: *IEEE Data Eng. Bull.* 33.2, pp. 32–39.

Zhong, M., Sharma, S. and Lingras, P. (2005). 'Refining genetically designed models for improved traffic prediction on rural roads'. In: *Transportation Planning and Technology* 28.3, pp. 213–236.

Zhou, C., Frankowski, D., Ludford, P., Shekhar, S. and Terveen, L. (2007). 'Discovering personally meaningful places'. In: *ACM Transactions on Information Systems* 25.3, 12–es.

Zhu, X., Li, J., Liu, Z., Wang, S. and Yang, F. (2016). 'Learning transportation annotated mobility profiles from GPS data for context-aware mobile services'. In: *Proceedings - 2016 IEEE International Conference on Services Computing, SCC 2016*, pp. 475–482.

Ziegler, J., Lategahn, H., Schreiber, M., Keller, C. G., Knoppel, C., Hipp, J., Haueis, M. and Stiller, C. (2014). 'Video based localization for bertha'. In: *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, pp. 1231–1238.

Zolliker, M., Rollier, R. and Bosshard, A. (2015). 'Co-creation eines Smart-Data-Werkzeuges zur Verkehrsmessung'. In: *4. nationale Smart City Tagung*.

Online references

Department of Commerce, U. (2017). *National Oceanic and Atmospheric Administration*. URL: https://www.ncdc.noaa.gov.

Der Schweizerische Bundesrat (2017). *Botschaft zum Bundesgesetz über die Totalrevision des Bundesgesetzes über den Datenschutz und die Änderung weiterer Erlasse zum Datenschutz*. URL: https://www.admin.ch/opc/de/federal-gazette/2017/6941.pdf.

Die Bundesversammlung der Schweizerischen Eidgenossenschaft (2017). *Bundesgesetz über die Totalrevision des Bundesgesetzes über den Datenschutz und die Änderung weiterer Erlasse zum Datenschutz*. URL: https://www.admin.ch/opc/de/federal-gazette/2017/7193.pdf.

Die Schweizerische Eidgenossenschaft (2018). *Federal Statistical Office*. URL: www.bfs.admin.ch (visited on 13/08/2018).

European Emergency Number Association (2018). *Advanced Mobile Location*. URL: http://www.eena.org/pages/aml (visited on 06/09/2018).

Federal Communications Commission (2018). *Enhanced 911 - Wireless Services*. URL: https://www.fcc.gov/general/enhanced-9-1-1-wireless-services (visited on 26/08/2018).

Federal Statistical Office (2018). *Mobile Internetnutzung*. URL: https://www.bfs.admin.ch/bfs/de/home/statistiken/kultur-medien-informationsgesellschaft-sport/informationsgesellschaft/gesamtindikatoren/haushalte-bevoelkerung/mobile-internetnutzung.html (visited on 10/07/2018).

Maldoff, G. (2016). *How GDPR changes the rules for research*. URL: https://iapp.org/news/a/how-gdpr-changes-the-rules-for-research/ (visited on 05/07/2018).

Newzoo (2018). *Top 50 Countries/Markets by Smartphone Users and Penetration*. URL: https://newzoo.com/insights/rankings/top-50-countries-by-smartphone-penetration-and-users/ (visited on 26/08/2018).

OpenStreetMap Contributors (2017). *OSM excerpt obtained via Geofabrik*. URL: https://www.openstreetmap.org (visited on 17/10/2017).

PeMS (2017). *Performance Measurement System*. URL: http://pems.dot.ca.gov/ (visited on 23/05/2017).

Swiss Confederation (2014). *Federal Act on Data Protection*. URL: https://www.admin.ch/opc/en/classified-compilation/19920153/index.html.

Swiss Federal Office of Transport (2018). *Open Data Platform Swiss Public Transport*. URL: https://opentransportdata.swiss/ (visited on 19/04/2018).

Swisscom (2018). *Smart City Montreux*. URL: https://www.swisscom.ch/de/business/enterprise/themen/digital-business/des-2017-007-smart-city-montreux.html (visited on 10/07/2018).

The Federal Authorities of the Swiss Confederation (2012). *Ordinance to the Federal Act of Data Protection*. URL: http://www.admin.ch/opc/de/classified-compilation/19930159/201210160000/235.11.pdf.

Curriculum Vitae

Oliver Burkhard,
born on 1987-01-20 in Thun (BE),
citizen of Schwarzhäusern (BE).

Education

2015–2018	Dissertation at the Geographic Information Systems Division in the Department of Geography, University of Zurich. Title of the thesis: *Towards Passive Tracking and Analyses of Human Mobility at Population Scale*, supervised by Prof. Dr. R. Weibel.
2011–2014	Qualifying for the title of Actuary SAA awarded by the Swiss Association of Actuaries.
2009–2011	MSc. Math. ETH at the Seminar for Statistics at the ETH Zurich. Title of the thesis: *The Effects of Managed Care Models on Health Care Expenditure*, supervised by Marloes Maathuis.
2009	Spring term at Lund University as an Erasmus exchange student.
2006–2009	BSc. Math UZH at the Institute for Mathematics at the University of Zurich.
2002–2005	Matura, emphasis on Physics and applications of Mathematics at the Gymnasium Thun Schadau. Content and language integrated learning in English.

Professional experience

2015–2018	Research assistant at the Department of Geography at the University of Zurich
2011–2016	Life Actuary at AXA Winterthur in Winterthur and Paris
2008 – 2011	Various undergraduate teaching activity at the University of Zurich and ETH Zurich.

Publications

Burkhard, O., Ahas, R., Saluveer, E. and Weibel, R. (2017). 'Extracting regular mobility patterns from sparse CDR data without a priori assumptions'. In: *Journal of Location Based services* 11.2, pp. 78–97.

Burkhard, O., Becker, H., Weibel, R., Axhausen, K.W. (in review). 'Minimal requirements on spatial accuracy and spmaling rate for transport mode detection in view of an imminent shift to passive signalling data'. In: *Computers, Environment and Urban Systems*.

Burkhard, O., Weibel, R. (in review) 'Improving the accuracy of deep learning based traffic flow predictions using robust features and a residual problem statement'. In: *IET Intelligent Transport Systems*.