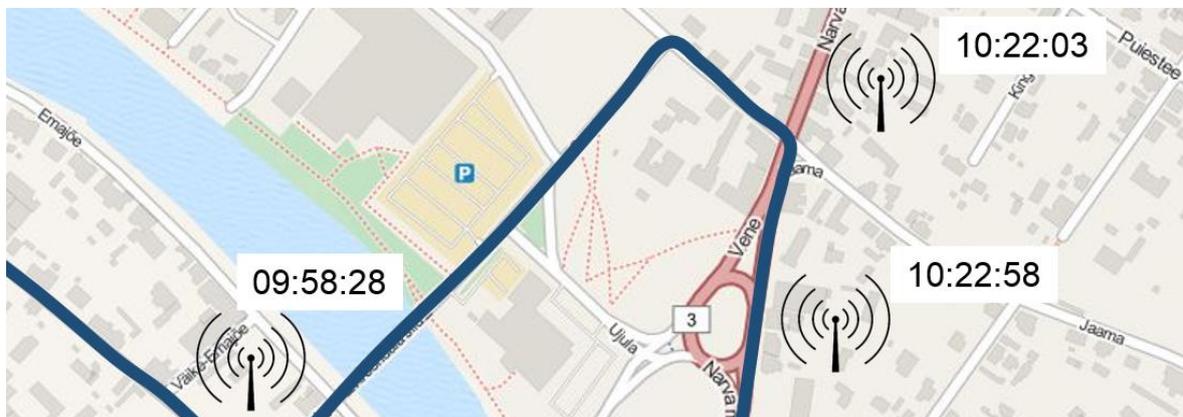


Master's Thesis GEO 511

Reconstructing Trajectories from Sparse Call Detail Records

A Case Study with Estonian Mobile Phone Data



Submitted by: Michelle Fillekes (08-730-103)

Date of Submission: June 30, 2014

Supervisors: Dr. Patrick Laube & Dr. Walied Othman

Faculty Member: Prof. Dr. Robert Weibel

Geographic Information Science (GIS)

Department of Geography
University of Zurich – Irchel
Winterthurerstrasse 190
8057 Zurich – Switzerland

External Supervisor: Erki Saluveer

erki.saluveer@positium.ee

Department of Geography

University of Tartu

Vanemuise St. 46

51014 Tartu – Estonia

Figure on Title Page

Source of underlying map: © OpenStreetMap (2014)

Contact

Author

Michelle Fillekes

Riedtlistrasse 70
8006 Zurich – Switzerland
michelle.fillekes@gmail.com

Supervisors

Dr. Patrick Laube

Geoinformatics Research Group
Institute of Natural Resource Sciences
ZHAW Zurich University of Applied Sciences
Grüental, Postfach
8820 Wädenswil – Switzerland
patrick.laube@zhaw.ch

Dr. Walied Othman

Geographic Information Science (GIS)
Department of Geography
University of Zurich – Irchel
Winterthurerstr. 190
8057 Zurich – Switzerland
walied.othman@geo.uzh.ch

Prof. Dr. Robert Weibel

Geographic Information Science (GIS)
Department of Geography
University of Zurich – Irchel
Winterthurerstr. 190
8057 Zurich – Switzerland
robert.weibel@geo.uzh.ch

Erki Saluveer

Department of Geography
University of Tartu
Vanemuise St. 46
51014 Tartu – Estonia
erki.saluveer@positium.ee

Acknowledgements

Now, approaching the end of my studies at the Geography Department of the University of Zurich, I would like to express my gratitude to all the people supporting me and contributing towards making this a wonderful, inspiring and instructive time.

For your support during the completion of this Master's thesis, I would like to express special thanks to:

- Dr. Patrick Laube and Dr. Walied Othman, for the fruitful discussions, constructive input, creative ideas and the constant encouragement in meetings within and outside the department.
- Erki Saluveer, for the valuable cooperation and the generous support, especially in all data-related questions.
- Susan McStea, for the very helpful linguistic revision; from now on I promise to never again confuse “number” and “amount”.
- My very dear friends and family and especially Jan Sennekamp for being there.
- And huge thanks to my parents, Iris and Jasper Fillekes, for their continuous support during my studies.

Thank you!

Michelle Fillekes
University of Zurich
June 2014

Abstract

For billing purposes, mobile phone operators collect large volumes of data containing information about time of phone activities (such as calls or SMS), as well as the location of the antennas people were connected to when initiating or receiving phone activities in so-called Call Detail Records (CDRs) (Ahas et al., 2008a). The Geography Department of the University of Tartu has a long-standing collaboration with Positium LBS (2014) which handles mobile positioning data from the Estonian mobile phone operators. For this thesis, the one month CDRs of 6 mobile phone users were provided as extracts of these data. Considerable research (e.g., Ahas et al., 2008a; Candia et al., 2008; Furletti et al., 2012) has been conducted analyzing CDR data that typically feature low spatial resolution (depending on the density of the antenna network) and are sampled in irregular temporal intervals (depending on the frequency of phone calls). Most of this work, however, focuses on the description of the spatial distribution of different human activities and not on the analysis of the actual movements expressed by individuals.

An aim of the thesis is to develop techniques to reconstruct people's trajectories in geographic space from the sparsely sampled CDR data. Subsequently, the trajectory reconstruction methods are validated by comparing the reconstructed trajectories to the GPS trajectories available for the same users having a much finer temporal and spatial granularity. A further aim is to investigate whether CDR data conditions (e.g., number of CDR fixes) under which a more accurate trajectory reconstruction is to be expected can be established.

In several pre-processing steps, the one month CDR and GPS data of the six users are divided into daily segments and then clipped to the time frame shared by the GPS and the CDR data. To reconstruct the paths from the CDR segments, in a first step, different methods to match the antenna locations to the most reasonable node on the OpenStreetMap (OSM) road network are applied. In a second step, the shortest path between the identified nodes on the road network is computed.

To validate the proposed methods, measures are computed that are aimed at assessing the similarity between the reconstructed and the corresponding GPS trajectories. The validation shows that overall, less than 30% of the actual travelled path can be reconstructed. The examination of the impact of the properties of the underlying CDR data on the accuracy of the reconstructed paths indicates that an increasing number of spatially unique CDR fixes and movements of an increasing length clearly have a positive impact on the accuracy of the trajectory reconstruction. The impact of the temporal resolution of the CDR data is marginal.

Zusammenfassung

Für Abrechnungszwecke sammeln Mobilfunkunternehmen in so genannten Call Detail Records (CDRs) grosse Mengen an Daten. Diese enthalten Informationen über die Uhrzeit von Telefonverbindungen (Anrufe, SMS, etc.) sowie die Position der dazu verwendeten Antennen (Ahas et al., 2008a). Das Geographische Institut der Universität Tartu pflegt eine langjährige Zusammenarbeit mit der Positium LBS (2014), welche Zugang zu den Mobilfunkdaten der estnischen Mobilfunkanbieter hat. Auszüge dieser Daten wurden für die vorliegende Masterarbeit zur Verfügung gestellt. Es existieren bereits viele Forschungsarbeiten (z.B., Ahas et al., 2008a; Candia et al., 2008; Furletti et al., 2012), welche sich mit der Analyse von CDR Daten beschäftigen. CDRs weisen eine geringe räumlich Auflösung, welche abhängig von der Dichte des Mobilfunknetzes ist, auf. Des Weiteren liegen diese Daten in zeitlich unregelmässigen Abständen, abhängig von der Häufigkeit von Telefonverbindungen, vor. Die meisten Analysen sind dabei darauf ausgerichtet die räumliche Verteilung von menschlichen Aktivitäten zu beschreiben und nur wenige Studien beschäftigen sich mit der Beschreibung von Bewegungen, welche von Individuen ausgeführt wurden.

Ein Ziel dieser Masterarbeit ist es, Methoden zu entwickeln, welche die Rekonstruktion menschlicher Trajektorien im geographischen Raum auf Basis von CDR Daten ermöglichen. Durch den Vergleich der rekonstruierten Trajektorien mit den entsprechenden, zeitlich als auch räumlich höher aufgelösten, GPS Trajektorien sollen die entwickelten Methoden validiert werden. Ein weiteres Ziel dieser Arbeit ist es, zu untersuchen, ob bestimmte Eigenschaften der CDR Daten (z.B. die Anzahl an Mobilfunkverbindungen) dazu beitragen, dass bessere Rekonstruktionsgenauigkeiten erreicht werden können.

In einem ersten Schritt werden dazu die CDR und GPS Daten, welche für sechs Mobilfunkteilnehmer über einen Zeitraum von jeweils einem Monat zur Verfügung stehen, in Segmente unterteilt, welche die Positionen von jeweils einem Tag enthalten. Um zu gewährleisten, dass die zusammengehörenden CDR und GPS Segmente jeweils die gleiche Zeitspanne abdecken, werden die Segmente entsprechend zugeschnitten. Darauf aufbauend soll versucht werden, anhand der CDR Segmente die zurückgelegten Trajektorien zu rekonstruieren. Dazu werden in einem ersten Schritt verschiedene Map-Matching Methoden angewendet, um ausgehend von den Standpunkten der Antennen die wahrscheinlichste Position auf dem OpenStreetMap (OSM) Strassennetzwerk zu ermitteln. In einem zweiten Schritt wird der kürzeste Weg zwischen den identifizierten Positionen auf dem Strassennetzwerk berechnet.

Zur Validierung der entwickelten Methoden, werden verschiedene Masse berechnet, welche den Zweck haben, die Ähnlichkeit zwischen einer rekonstruierten und einer GPS Trajektorie zu erfassen. Die Analyse der Ähnlichkeitsmasse ergibt, dass insgesamt weniger als 30% der tatsächlich zurückgelegten Wege rekonstruiert werden können. Eine folgende Evaluierung bezüglich der Auswirkung verschiedener Eigenschaften der zugrundeliegenden CDR Daten auf die Genauigkeit der rekonstruierten Wege weist darauf hin, dass eine zunehmende Anzahl von CDR Positionen sowie eine zunehmende Länge der gesamten Bewegung einen eindeutig positiven Effekt auf die Qualität der rekonstruierten Trajektorien haben. Der Einfluss der zeitlichen Auflösung hingegen ist minimal.

Contents

Acknowledgements	I
Abstract	III
Zusammenfassung	V
Contents	VII
List of Figures	IX
List of Tables	XI
List of Abbreviations	XIII
1 Introduction	1
1.1 Context	1
1.2 Objectives and research questions	1
1.3 Thesis structure	2
2 Related Work	3
2.1 Human movement tracking with mobile positioning techniques	3
2.1.1 Mobile positioning techniques	3
2.1.2 Application areas of mobile positioning data	5
2.1.3 Privacy issues and data security	5
2.2 Studies with CDR data	6
2.2.1 Studies of spatio-temporal patterns of human activities	6
2.2.2 Studies regarding mobility of mobile phone users	7
2.3 Algorithms relevant for trajectory reconstruction	9
2.3.1 Map matching	9
2.3.2 Centrality measures in road networks	9
2.3.3 Automatic route selection	10
2.4 Assessment of trajectory similarity	11
2.4.1 Spatial similarity measures	11
2.4.2 Spatio-temporal similarity measures	13
2.5 Identification of research gaps and research questions	14
3 Overview of the overall workflow and the software used	17
4 Data and pre-processing	19
4.1 Overview	19
4.2 Positioning data	19
4.2.1 CDR data	20
4.2.2 GPS data	22
4.3 Pre-processing of positioning data	23
4.3.1 Time zone conversions and coordinate system transformation	23
4.3.2 Segmenting the CDR data	23
4.3.3 Clipping the CDR and GPS daily segments for similar time frames	25
4.3.4 Excluding daily segments unsuitable for reconstruction	25
4.4 OSM road network data	27

4.5	Pre-processing of OSM road network data	28
4.5.1	Making OSM road network routable.....	28
4.5.2	Transforming OSM road network to a GeoTools graph model	29
4.6	Possible CDR, GPS, OSM data constellation	30
5	Trajectory reconstruction based on CDR data	31
5.1	Overview	31
5.2	Matching CDR data to a node on the network.....	31
5.2.1	MM method 1: Approach relying on proximity of nodes to antenna	32
5.2.2	Computing the Voronoi diagram	33
5.2.3	MM method 2: Approach relying on center of gravity of Voronoi cells	34
5.2.4	MM method 3 and 4: Approaches relying on degree centrality of nodes ..	35
5.2.5	MM methods 5-7: Approaches relying on edge-based criteria.....	36
5.2.6	Pre-validation of CDR map matching	38
5.3	Shortest path between the selected nodes	39
5.4	Summary of trajectory reconstruction methods	41
6	Validation.....	43
6.1	Overview	43
6.2	Making paths from GPS fixes	43
6.2.1	1 st step: Identifying edges closest to GPS fixes.....	44
6.2.2	2 nd step: Removing unintentionally identified edges	45
6.2.3	3 rd step: Making GPS path continuous	46
6.2.4	Disqualification of unsuitable ground truth paths.....	46
6.3	Assessing similarity between reconstructed and ground truth paths	47
6.3.1	Units to be compared.....	47
6.3.2	Computation of similarity measures	48
6.3.3	Discussion of similarity measures	51
6.3.4	Selection of similarity measures.....	52
6.4	Results	53
6.4.1	Comparison of different trajectory reconstruction algorithms.....	53
6.4.2	Impact of CDR data properties on accuracy of trajectory reconstruction ..	56
7	Discussion.....	65
7.1	Methods to reconstruct trajectories from sparse CDR data	65
7.2	Validation of the trajectory reconstruction methods.....	68
7.3	Impact of CDR data properties on trajectory reconstruction accuracy	70
8	Conclusion	73
8.1	Summary	73
8.2	Contributions	73
8.3	Outlook	74
9	References	75
	Personal declaration	85

List of Figures

Figure 1: Schematic representation of similarity measure used by Newson and Krumm (2009).....	12
Figure 2: Locality in-between polylines (Pelekis et al., 2011)	13
Figure 3: Workflow of the pre-processing of the positioning data	19
Figure 4: Workflow of pre-processing of the OSM road network	19
Figure 5: Average phone activity dynamics (Csáji et al., 2013)	24
Figure 6: Principle of the clipping of the daily segments: fixes crossed out in grey are discarded due to lack of corresponding data of the opposite data source.....	25
Figure 7: Frequency diagram representing the number of gaps as a function of their lengths.....	26
Figure 8: Diagram representing number of daily segments to exclude as a function of the tolerated maximum gap length between consecutive GPS fixes	27
Figure 9: OSM road network of Estonia, showing the added and removed lines.....	29
Figure 10: Initial OSM, CDR, GPS constellation (subset of CDR / GPS data of user 4, 03.06.2013)	30
Figure 11: Workflow of trajectory reconstruction	31
Figure 12: Node on the road network closest to the CDR fix (subset of CDR data of user 4, 03.06.2013).....	32
Figure 13: Voronoi cells describing the area closest to each antenna (Estonian mobile phone operator's antennas, partially slightly shifted, June 2013).....	33
Figure 14: Center of gravity of the Voronoi cells used (subset of CDR data of user 4, 03.06.2013)	34
Figure 15: Highest one-level and two-level degree centrality (subset of CDR data of user 4, 03.06.2013).....	35
Figure 16: Edges identified on edge-based criteria: the two criteria highest road category and maximum speed identify the same edge, the respective road category (secondary) and the maximum speed (50km/h) are indicated (subset of CDR data of user 4, 03.06.2013).....	37
Figure 17: Box plots showing the Euclidean distances between the map-matched CDR fixes (according to MM methods 1-7) and the temporally closest GPS fixes, the diamond represents the mean.....	39
Figure 18: Reconstructed trajectories on the basis of MM method 6 (TR 6, maximum speed rationale) and MM method 7 (TR 7, longest edge rationale) (subset of CDR data of user 4, 03.06.2013).....	41
Figure 19: Workflow of validation of trajectory reconstruction methods by comparing reconstructed paths to the ground truth	43

Figure 20: Visualization of initial GPS fixes and the three steps for deriving the continuous GPS path (subset of GPS data of user 4, 03.06.2013)	44
Figure 21: Distance calculation between a point and a line segment defined by A^0 and A^1 (White et al., 2000)	45
Figure 22: Diagram of average distance from an edge to the closest GPS fix per ground truth path	47
Figure 23: Path reconstructed by TR method 3 and corresponding ground truth path, allowing quantification of the similarity of the two paths (based on CDR and GPS data of user 4, 03.06.2013)	47
Figure 24: Scatter plot matrix for the different similarity measures including correlation values for all the 511 comparison cases	51
Figure 25: Scatter plot for SMs 4 and 12 for all 511 comparison cases, color-coded according to the TR method	54
Figure 26: Box plots of the values of SMs 4 and 12 for the 73 reconstructed daily paths with the different TR methods	55
Figure 27: Path reconstructed by TR method 6 and corresponding ground truth path, plus associated similarity measures and statistical properties of the data used (based on CDR and GPS data of user 4, 11.06.2013)	56
Figure 28: Path reconstructed by TR method 6 and corresponding ground truth path, plus associated similarity measures and statistical properties of the data used (based on CDR and GPS data of user 6, 22.06.2013)	57
Figure 29: Path reconstructed by TR method 6 and corresponding ground truth path, plus associated similarity measures and statistical properties of the data used (based on CDR and GPS data of user 5, 28.06.2013)	57
Figure 30: Bar chart for TR method 6 with SMs 4 and 12 resulting from an augmentation of the minimum number of required CDR fixes with unique locations, thereby reducing the number of daily segments considered.....	59
Figure 31: Scatter plot for SMs 4 and 12, color-coded according to two classes of number of spatially unique CDR fixes used as input	59
Figure 32: Bar chart representing the avg. SM for reconstructed paths with TR method 6 for three groups of avg. time between consecutive CDR fixes	61
Figure 33: Scatter plot for SMs 4 and 12 of all 511 comparison cases, color-coded according to three classes of avg. time between consecutive CDR fixes.....	61
Figure 34: Bar chart with average SMs 4 and 12 for daily segments grouped into three classes of distance of movement.....	63
Figure 35: Scatter plot representing the correlation between SMs 4 and 12 for all 511 comparison cases, color-coded according to three distance of movement classes	63

List of Tables

Table 1:	Initial number of CDR and GPS fixes per user	20
Table 2:	Excerpt from the original CDR data of user 6.....	20
Table 3:	Summary of the most important features of the CDR data	22
Table 4:	Excerpt from the original GPS data of user 1	22
Table 5:	Summary of the most important features of the GPS data	23
Table 6:	Statistics for daily CDR segments per user.....	24
Table 7:	Attributes of routable road features.....	28
Table 8:	Assigning a ranking to road types	36
Table 9:	Descriptive statistics for the distances in m between CDR and GPS fixes for the MM methods 1-7	39
Table 10:	Overview of trajectory reconstruction (TR) methods and the respective underlying map-matching (MM) methods.....	41
Table 11:	Overview of the different proposed and implemented similarity measures (SMs).....	48
Table 12:	Descriptive statistics for SM 4 for the different trajectory reconstruction methods.....	55
Table 13:	Descriptive statistics of SM 12 for the different trajectory reconstruction methods.....	55
Table 14:	Class boundaries for the three equal-sized groups of low, medium and high temporal resolution according to the average time difference bet- ween consecutive CDR fixes	61
Table 15:	Class boundaries for the three equal-sized distance groups according to the path length derived from the GPS fixes.....	62

List of Abbreviations

AOA	Angle of Arrival
BTS	Base Transceiver Station
CDR	Call Detail Record
CE(S)T	Central European (Summer) Time
CGI	Cell Global Identity
DSPF	Dijkstra Shortest Path Finder
EE(S)T	Eastern European (Summer) Time
GIS	Geographic Information System
GPS	Global Positioning System
GSM	Global System of Mobile Positioning
HMM	Hidden Markov Model
IDE	Integrated Development Environment
ITS	Intelligent Transportation System
LA	Location Area
LAC	Location Area Code
LBS	Location Based Services
LIP	Locality In-between Polylines
MM	Map Matching
RSSI	Received Signal Strength Indicator
SM	Similarity Measure
TA	Timing Advance
TOA	Time of Arrival
TR	Trajectory Reconstruction
UTC	Coordinated Universal Time
UTM	Universal Transverse Mercator
VLR	Visitor Location Register
WGS 84	World Geodetic System 1984

1 Introduction

1.1 Context

For billing and network management purposes, mobile phone operators collect large amounts of data containing information regarding time of phone activities (calls, SMS, etc.) as well as the location of the antenna that routed the respective phone activities in so-called Call Detail Records (CDRs) (Ahas et al., 2008a). These data are of interest because they can reveal information about the activities of a large number of people in space and time. The spatial resolution of the CDRs is dependent on the density of the mobile phone network as well as the coverage area of an antenna (Caceres et al., 2007). The temporal resolution is irregular as it depends on the number and time intervals of phone activities. The advantage of CDR data is that mobile phone use is very popular (mobile phone penetration of 95% in Estonia) and therefore information about the behavior of a large fraction of the population over long time periods can be acquired by studying this data source (TNS, 2014). This kind of data, however, is generally not easily obtainable from mobile operators and the use of it has important implications for the privacy rights of the people concerned. For this study, the CDR and the corresponding GPS data of 6 mobile phone users over a one month period are provided by Positium LBS (2014) in collaboration with its long-standing partner in academia, the Geography Department of the University of Tartu and consent to use the data was obtained from mobile phone users involved in this study.

Substantial research efforts have been expended on the analysis of CDR data. Most studies investigate the spatial distribution of a human activity, e.g., the seasonality of foreign tourists' space consumption in Estonia by Ahas et al. (2007a). The less prevalent studies that focus on the mobility of individuals often focus on the identification of proxies, e.g., radius of gyration, in order to describe a typical range of a mobile phone user (Blumenstock, 2012; Csáji et al., 2013; Frias-Martinez et al., 2010). The few research attempts (e.g., Doyle et al., 2011; Wang et al., 2010) that try to reconstruct individuals' movements based on CDR data only function in constrained settings (e.g., between a specific origin and destination). The study of human mobility is of particular interest with respect to, e.g., traffic prediction systems or urban planning (e.g., Brakatsoulas et al. 2005). Saluveer and Ahas (2014) recognize considerable potential for mobility studies based on CDR data.

1.2 Objectives and research questions

Prediction of people's movement behavior on the basis of CDR data, which typically feature low temporal and spatial resolutions, requires development of new methods based on various assumptions. An aim of this thesis is to propose methods to reconstruct individuals' trajectories in geographic space from CDR data. In order to validate the proposed methods, the reconstructed trajectories are compared to the actively tracked GPS trajectories of the same journeys having a much finer temporal and spatial granularity. A last aim of this thesis is to examine whether CDR data properties (e.g., number of fixes) have an impact on the quality of the reconstructed trajectories. To address the above mentioned aims the following research questions (RQ) will be examined:

RQ 1: How can mobile phone users' trajectories be reconstructed from sparsely sampled CDR data?

RQ 2: In order to validate the trajectory reconstruction methods developed in this study, what level of similarity can be achieved by comparison of the reconstructed trajectories with higher resolution GPS trajectories of the same journeys?

RQ 3: Which properties of the CDR data, such as sampling properties or trajectory length, affect the accuracy of the reconstructed trajectories?

1.3 Thesis structure

Following the introduction, Chapter 2 provides an overview of work related to trajectory reconstruction and mobile positioning data. Chapter 3 gives an overview of the methodology used in the thesis. Chapter 4 presents the data and the pre-processing steps required for the trajectory reconstruction and validation. Chapter 5 describes the methods developed to reconstruct trajectories based on CDR data. Chapter 6 explains the validation of the trajectory reconstruction methods and discusses the results. In Chapter 7 the research questions are discussed and the results are placed into the context of the related work. Finally, Chapter 8 summarizes the most important aspects of the thesis; furthermore, the main contributions are underlined and an outlook for future work is given.

2 Related Work

This chapter gives an overview of related work and background information of concepts and algorithms that are used in this thesis. Particularly, the following topics will be considered: overview of different human movement tracking techniques based on mobile phones, their application areas and the involved privacy issues (Section 2.1); a selection of studies using CDR data regarding the geography of human activities and the mobility of the users (Section 2.2); existing algorithms and relevant concepts to reconstruct trajectories such as map matching, centrality measures and automatic route selection (Section 2.3); and concepts to assess similarity of trajectories that are used for the validation of the algorithms (Section 2.4). In Section 2.5, the research gaps are identified and the research questions of this thesis are presented.

2.1 Human movement tracking with mobile positioning techniques

Human movement tracking is becoming increasingly more important because many human daily activities such as payment with a credit card, the use of a mobile phone or the use of other location-aware devices allow the inference of a person's current geographic location (Giannotti et al., 2007). The analysis of such locational information may contribute to a better understanding of how society works. Many modern mobile phones possess GPS receivers, which are explicitly designed to accurately assess a person's position. In this section, however, the focus is on mobile positioning techniques through cellular networks – generally featuring a lower spatial and temporal resolution but potentially being available for a large fraction of the population over long time periods – as well as their possible application areas and issues regarding privacy and data security. Cellular network-based techniques are to be categorized amongst the Eulerian sensing methods, which collect measurements at pre-defined points (Work et al., 2009). This is in contrast to Lagrangian measurements that are performed by a sensor (e.g., GPS device) moving along a trajectory.

2.1.1 Mobile positioning techniques

Especially within the domain of location based services (LBS), research was driven in the direction of identifying a mobile phone user's position within a cellular network (Mountain and Raper, 2001a; Ratti et al., 2006). Smoreda et al. (2013) introduce the Global System for Mobile Communication (GSM) cellular network as a radio network of base transceiver stations (BTSs), each of which having one or more antennas. The BTSs are distributed in a manner that allows best possible radio coverage via small regions called cells. Devices such as mobile phones can access the phone network via the BTSs. Therefore the mobile phone's position is identified using the cell global identity (CGI). Subsequently, the CGI can be matched with the coordinates or the cell coverage of the respective BTS station, which again can be used as approximate location of a mobile phone. When a mobile phone user moves from one cell into another during a phone call, the antenna is automatically switched. The antenna switches are designated as handovers (Zuo et al., 2012). In order to manage the user's mobility, the cells are aggregated into bigger subdivisions of the cellular network,

referred to as location areas (LA). The location area code (LAC) gives information about the LA in which a user is located which enables the network to optimize the network traffic. In so-called visitor location registers (VLRs) the LAC and the last known CGI are stored. The VLR is updated each time a user moves from one LA to another, when the handset is switched on or off, or after a longer period of several hours with no mobile phone activity (Caceres et al., 2007; Saluveer and Ahas, 2014; Smoreda et al., 2013).

Depending on the purpose for which the mobile phone data are collected or used, selected components of the above-mentioned information of the cellular network are stored or can be accessed. Additional information sources relevant for positioning consist of, inter alia, the received signal strength indicator (RSSI) and the timing advance (TA). As described in Waadt et al. (2009), the RSSI indicates the signal strength which is a function of the distance between BTS and the mobile device and can consequently be used as an approximation for the distance between the BTS and the handset. The TA gives an indication of the signal propagation delay from the BTS to the handset and back, which also can be transformed into a distance estimation between handset and BTS. If the time lag between handset and BTS for three or more BTSs are available, the mobile user's location can be further narrowed down to the overlapping area of the respective coverage areas, as described in the time of arrival (TOA) technique in Ahas and Laineste (2006). The angle of arrival (AOA) estimates the user's position based on the angle at which the signals from at least two BTSs arrive (Zang et al., 2010).

Call detail records (CDRs), which are at the center of this thesis, are archived for billing purposes or technical network management by mobile phone operators. CDR data usually comprise at least the following attributes: time of phone activity, CGI (which can be assigned to the antenna's¹ coordinates) and user ID (Ahas et al., 2010b; Järv et al., 2012). Depending on the mobile phone operator, different in- and / or outgoing mobile phone activities such as SMS, MMS, calls, and / or internet connections are included. Typically, CDR data including internet connections (also referred to as data detail records) have a higher temporal resolution, since many services on a mobile phone regularly connect to the internet (Saluveer and Ahas, 2014). Optionally, further attributes such as call duration, the ID of the call / message receiver and cell handovers during a phone activity are contained in the CDRs (Bar-Gera, 2007; Csáji et al., 2013).

CDRs are also referred to as passive mobile positioning data, thereby alluding to the fact that it is automatically stored in the log files of mobile service providers (Ahas et al., 2009, 2008a; Smoreda et al., 2013; Toomet et al., 2011). Active mobile positioning data (also referred to as mobile tracing data), on the contrary, are collected after a special request to determine a mobile phone user's location (Ahas et al., 2010a). Thereby the temporal resolution of the data can be controlled, by requesting the mobile phone's location in regular temporal intervals.

¹ In the following, BTSs will consistently be referred to as antennas.

2.1.2 Application areas of mobile positioning data

The mobile phone penetration in developed countries is close to 100% (MDGS, 2014), therefore, knowledge of the behavior of almost the total population can potentially be extracted. Besides the advantage of the high penetration, CDR data is popular in research due to its relative ease of extraction (they are pre-processed by mobile phone operators in a standard format and recorded in highly secured databases) and they are available in huge quantities over long time periods (Smoreda et al., 2013). The following list comprises a non-exhaustive overview of important areas where CDR data (as well as other mobile positioning data) are applied:

- Location based services (LBSs), becoming increasingly important with the spread of smart phones (Asakura and Hato, 2004; Mountain and Raper, 2001a, 2001b; Ratti et al., 2006; Raubal et al., 2004)
- Intelligent transportation systems (ITSs) for traffic prediction and management (Herring et al., 2010; Work et al., 2009)
- Transportation infrastructure and public transportation planning (Saluveer and Ahas, 2014)
- Regional, urban / spatial planning (Ahas et al., 2008a; R. Ahas et al., 2007; González et al., 2008)
- Identifying important tourism destinations and tourist tracking for management and promoting purposes (Rein Ahas et al., 2007a; Andres et al., 2009; Shoval and Isaacson, 2007)
- Trend analyses or space-time variability studies including migration studies (Ahas et al., 2008a; Blumenstock, 2012)
- Surveillance and alibi queries for security or military services (de Montjoye et al., 2013; Gudmundsson et al., 2008; Kuijpers et al., 2010; Michael et al., 2006)
- Studies of social interactions and structure (Candia et al., 2008; Palla et al., 2007; Phithakkitnukoon et al., 2010; Winter and Kealy, 2012)

2.1.3 Privacy issues and data security

Despite the important benefits that (mobile) positioning data represent for many application areas, there are major concerns regarding privacy issues and data security. As stated in Michael et al. (2006), humans are mostly unwillingly tracked. One should be aware that positioning data can be misleading. Wrong interpretations of an individual's behavior might put the person in a bad light, which might unfairly damage his reputation. De Montjoye et al. (2013) reveal in their study how easily people are identifiable from temporally and spatially very coarse data.

Many authors are aware of privacy implications when using positioning data. Most of them come to the conclusion, however, that this data source brings far more advantages – when used in a controlled setting – than it brings harm (Michael et al., 2006; Ratti et al., 2006). According to a survey led by Ahas et al. (2010a) only 10% of the participants would not participate in a tracking study due to fear of surveillance. Fear of surveillance is dependent

on the cultural and the societal background, but is also a personal matter. Therefore, individuals should have the possibility to refuse being tracked and this without being considered as suspicious.

Many authors (Ahas et al., 2010b; Ratti et al., 2006; Saluveer and Ahas, 2014) refer to the *Directive on Privacy and Electronic Communications* of the European Parliament (2002), which provides regulations concerning location data. According to these regulations, location data need to be received and treated in aggregated and anonymous form, in order that the linkage of location data with real people is not possible. Otherwise the consent of the users to the extent and the duration necessary has to be expressed (European Parliament, 2002). As a consequence, researchers apply various approaches in order to respect data security and thereby guarantee privacy of participants. For example, Smoreda et al. (2013) aggregate their data and investigate only small temporal units, whereas Doyle et al. (2011) delete the user IDs in their study.

2.2 Studies with CDR data

A considerable number of papers using mobile phone – specifically CDR – data have already been published. As mentioned in Section 2.1.1, CDR data differ in temporal resolution depending on inclusion of in- and / or outgoing phone activities, on the different kinds of included activities (SMS, MMS, calls, internet connections, etc.), as well as on whether handovers during phone calls are recorded or not. The component which is identical for all types of CDR data is the spatial resolution, which is equal to the coverage area or the location of an antenna. Studies using primarily active mobile positioning data (Ahas et al., 2010a; Andrienko et al., 2010; Ratti et al., 2006), which assess the mobile phone users' positions in regular temporal intervals and therefore typically feature a higher temporal resolution, are not considered in the following two sections.

Csáji et al. (2013) differentiate between the following three predominant research areas where mobile phone data are used: social structure, temporal dynamics and mobile behavior of mobile phone users. Mobile phone activities used as proxy for the structure and dynamics of social networks mainly rely on the information of (the number of) phone activities taking place between different phone users, whereas typically users are modeled as nodes and the relations between them as links of different importance depending on the mobile phone interactions between the respective users (Eagle et al., 2009b; Hidalgo and Rodriguez-Sickert, 2008; Onnela et al., 2007; Palla et al., 2007). Since the actual geographical location is mostly not taken into account in these studies, they are not further examined in the following two sections. Studies within the two remaining – slightly adapted – categories of Csáji et al. (2013) are presented the Sections 2.2.1 and 2.2.2,.

2.2.1 Studies of spatio-temporal patterns of human activities

The selection of studies presented in this section mostly aims at describing the spatial distribution of an investigated phenomenon regarding human behavior on an aggregated level, and, optionally, its temporal dynamics. Frequently, the information derived from these studies can be used for spatial planning purposes.

Candia et al. (2008) describe spatio-temporal calling patterns in an urban area with the aim of being able to detect anomalous events – such as an emergency situation – which would exhibit different patterns. Eagle et al. (2009a) show that differences in capital, urban and rural areas regarding average amount of monthly travel, average number of outgoing phone activities, or average tie strengths between subscribers exist. An approach called activity-aware mapping introduced by Phithakkitnukoon et al. (2010) tries to capture differences in daily activity patterns (e.g., working, shopping) between people who share different work-area profiles (basically location of workplace).

Also the group of Rein Ahas at The University of Tartu (Estonia) has conducted multiple studies with the aim of describing spatio(-temporal) patterns of human activities based on CDR data. Ahas et al. (2007a), for example, analyze the seasonality of foreign tourists' space consumption in Estonia. The method used therefore is called the social positioning method, which combines the use of locations (derived from the CDR data) and characteristics of the users (such as nationality) to describe tourists' space consumption. The objective is to use their results to implement regional planning measures. A model to predict space-time behavior of travelers as a function of air temperature is developed in Ahas et al. (2008a). Another paper of Ahas et al. (2010b) proposes an approach to determine locations meaningful to mobile phone users. Thereby, anchor points for all mobile phone users are computed in monthly intervals. According to the methodology of activity spaces, anchor points are classified into activity groups like home location or work location.

2.2.2 Studies regarding mobility of mobile phone users

The studies presented in this section have as primary aim the description of the mobile behavior of the mobile phone users. Many of the studies coming from the intelligent transportation system (ITS) research use positioning data with higher spatial (RSSI, TA, GPS data, etc.) and temporal resolutions (active mobile positioning, GPS data) and are therefore not further explored in this section (Asakura and Hato, 2004; Calabrese et al., 2011; Velaga et al., 2009; Waadt et al., 2009; Zuo et al., 2012). A good overview regarding methodological issues when using CDR data for mobility studies is given in Saluveer and Ahas (2014).

Wang et al. (2010) infer the transportation mode (e.g., walking, driving cars, public transport) of mobile phone users travelling between the same origin and destination based on travel times. Doyle et al. (2011) propose an alternative approach to the travel mode detection of Wang et al., who are only able to distinguish between travel modes that feature significantly different speeds. Therefore, the virtual cell paths for the two alternatives of a railway and a road are computed between a pre-defined source and a destination location. Subsequently, the number of cells that served phone activities of a specific user corresponding to either of the two virtual cell paths determine the more probable transportation mode alternative.

Through investigation of the interplay between human mobility and social ties, Wang et al. (2011) constitute one of the few exceptions amongst the social structure researchers working with mobile phone data who use geographical measures of closeness besides network closeness measures. Human mobility is thereby described by measures such as probability that users visit the same location (not necessarily) at the same time. Also Yuan et al. (2012)

make use of proxies to describe mobility of users, notably: the average lengths of the semi-major and major axes of an ellipsoid representing the probability of a person to be found at a certain location as depicted in Figure 3 in González et al. (2008, p. 782); the movement eccentricity representing the deviation of an ellipsoid from a circle, and the movement entropy, which characterizes the heterogeneity of a trajectory pattern based on the number and frequency of visited locations. The last measure was actually first used by Song et al. (2010) in the context of passive mobile phone data with the aim of defining the limits of predictability of human movements.

Likewise González et al. (2008) address the question of predictability of human movements by using several proxies for mobility such as the radius of gyration. This variable is a popular one used as approximation for users' mobility (Blumenstock, 2012; Csáji et al., 2013; Frias-Martinez et al., 2010; González et al., 2008; Song et al., 2010; Yuan et al., 2012). The radius of gyration is obtained by computing the root mean square distance of all phone activity positions from their center of mass (Frias-Martinez et al., 2010). This variable is used in literature to describe a typical range for a user's area of influence. Just as the variable number of different antennas that route phone activities of a mobile phone user is used as a proxy for a user's area of influence. Similarly, Csáji et al. (2013) and Blumenstock (2012) compute proxies, comparable to the above-mentioned, to describe users' mobile behavior. Additionally, the diameter of the convex hull devised from the CDR locations and the total length of the line segment resulting from a connection by straight lines of the CDR locations are used.

Smoreda et al. (2013) devise daily mobility motifs (particular sequence of a particular number of antennas visited) for each user on the basis of CDR data that start and end at the location that is identified as home location. The most prominent motif is the movement between two distinctive antennas (e.g., home and workplace). Smoreda et al. find that seven motifs describe over 80 % of the population's mobility motifs. Bar-Gera (2007) develops a method in which he makes use of registered handovers during a phone activity for a specific road section in order to devise speed indications for the road sections at different times. Based on the time, duration and the location of phone activities Furletti et al. (2012) differentiate between the following four categories of mobile phone users in a particular city: commuters, residents, tourists and people in transit.

Also recent work by the group of Rein Ahas focuses on the use of CDR data for mobility studies in Estonia (Järv et al., 2014, 2012; Saluveer and Ahas, 2014). Järv et al. (2012) propose a methodological approach that is able to distinguish between home and workplace commuting and other non-commuting movement based on phone activities along a specific road section. Thereby, the shortest-path heuristic is used to compute the connection between the home and workplace of the users crossing the investigated road segment. If the shortest paths of the respective users include the investigated road segment, it is classified amongst the commuting-related movements, otherwise amongst the non-commuting movements. Järv et al. (2014) approximate mobility by the number of unique mobile activity locations and thereby extract the 10 most-frequent activity locations. The variation of these numbers is further analyzed for a one-year period where seasonal effects on human travel behavior can be observed.

2.3 Algorithms relevant for trajectory reconstruction

This section provides an overview of the most important concepts that are used in the trajectory reconstruction algorithms developed in this study. In the first part, a selection of work from the map-matching research domain is presented (Section 2.3.1). In the second part, a short introduction to centrality measures is given (Section 2.3.2), and, in the final part, automatic route selection algorithms are presented (Section 2.3.3).

2.3.1 Map matching

Chawathe (2007, p. 1190) defines the map-matching problem as follows: “*The map-matching problem is, in general, the problem of correlating the path of a vehicle to a vector map of roads or other features.*” The problem thereby differs depending on the data source input (GPS, mobile positioning data, etc.) and whether the focus is on instantaneous positioning of vehicles (online map matching) or offline map matching that matches positions sampled in the past to a road network (Yin and Wolfson, 2004). The focus in this section is on offline map matching.

In literature, the following types of map-matching algorithms are typically distinguished: geometric, topological, probabilistic and advanced (Lou et al., 2009; Quddus et al., 2007). The simplest form of a geometric algorithm matches a position to the closest road node or edge (referred to as point-to-point and point-to-arc map matching, respectively). These approaches are sensitive to measurement noise which results in identification of wrong roads especially in dense urban road networks, where roads lie close together (Newson and Krumm, 2009). A more advanced geometric map-matching algorithm is proposed by White et al. (2000) where a geometric curve-to-curve matching algorithm is used, in which piecewise linear curves between the original fixes, as well as between candidate map-matched fixes are constructed and compared to each other in terms of distance.

A topological map-matching algorithm additionally makes use of contiguity of the road segments. Quddus et al. (2003) developed an enhanced topological map-matching algorithm which compares each point that is matched to an edge with the edge to which the previous point was matched. Probabilistic map-matching algorithms rely on elliptical or rectangular confidence regions surrounding the original positions (Quddus et al., 2007). These regions are derived, inter alia, from the error variances associated with the positioning device. Advanced map-matching algorithms use more refined concepts such as a Kalman Filter or the Hidden Markov model (HMM) (Krumm et al., 2007; Newson and Krumm, 2009; Quddus et al., 2007; Rahmani and Koutsopoulos, 2013). For example, Newson and Krumm (2009) apply a map-matching algorithm that uses a HMM and thereby place particular emphasis on data that are geometrically noisy and temporally sparse. Besides GPS data, positioning data from WiFi systems and cell tower multilateration are used.

2.3.2 Centrality measures in road networks

Centrality measures applied to the characterization of networks are used in various research domains such as sociology, biology or technology (Dorogovtsev and Mendes, 2001; Onnela et al., 2007; Reka and Barabási, 2002). Thereby, a network is typically represented

as a graph consisting of nodes connected by edges. In the context of road networks centrality measures are used, *inter alia*, for the following purposes:

1. to identify the most important / significant locations (nodes) on a road network (Crucitti et al., 2006),
2. characterization of road networks – e.g., self-organized vs. planned (Crucitti et al., 2008, 2006; Porta et al., 2006),
3. map generalization based on the centrality measure of a specific node / edge (Jiang and Claramunt, 2004; Jiang and Harrie, 2004).

A good overview of different centrality measures in the context of road networks is provided in Crucitti et al. (2006) or in Latora and Marchiori (2007). In the following, some commonly used centrality measures are presented. The degree centrality is based on the assumption that a central node is connected to a high number of adjacent nodes. Therefore, the degree centrality can be given, for example, by the number of adjacent edges of the node under consideration. The closeness centrality assesses how near a node is to all the other nodes. Crucitti et al. (2006) compute the measure by adding up the lengths of all the shortest paths from the investigated node to all the other nodes in the graph. The betweenness centrality in Crucitti et al. (2006) is defined by the number of shortest paths between all origin-destination combinations of a graph that traverse the investigated node. In order to measure information centrality, Latora and Marchiori (2007) propose to measure the relative drop of efficiency of a graph as a consequence of the removal of an investigated node (2007).

Centrality measures can be grouped according to some of their features. Crucitti et al. (2008) differentiate between topological centrality measures, which basically capture the number of steps (e.g., number of edges that are traversed) and spatial centrality measures, which consider metric distances (e.g., length of edges that are traversed). Furthermore, centrality measures can be categorized in local and global measures (Jiang and Claramunt, 2004; Jiang and Harrie, 2004). A local centrality measure considers the immediate neighborhood of an investigated node (e.g., degree centrality), whereas a global centrality measure considers the relationship of one node to all the other nodes of a graph (e.g., closeness centrality).

2.3.3 Automatic route selection

Automatic route selection between a pre-defined origin and destination on a road network is an extensively studied research subject. Thereby, the most commonly used approach is certainly the shortest-path algorithm, which has been studied for more than 40 years in diverse fields such as transportation and computer science (Fu et al., 2006).

For the shortest-path computation, edges of a road network are assigned weights which represent the costs involved in a traversal of the respective edges. For example, an edge is weighted according to its length or the average edge travel time (depending on maximum speed tolerated and edge length) (Yin and Wolfson, 2004). A shortest-path algorithm finds a path between an origin and a destination node that minimizes the sum of the costs of all the edges composing the respective path (Fu et al., 2006). If a graph is directed, which means that travel directions on each edge are defined, the shortest path is not necessarily

the same when source and destination node are reversed (Dorogovtsev and Mendes, 2001). Zhan (1997) differentiates between one-to-one (from one source node to all other nodes), one-to-some (from one source node to a subset of destination nodes) and all-to-all (from every node to every other node in the road network) shortest-path algorithms. A good overview of shortest-path algorithms is provided in Fu et al. (2006) or in Zhan (1997). One of the first algorithm in shortest-path computation is Dijkstra's shortest-path algorithm, named after its inventor (Dijkstra, 1959). The more efficient A* shortest-path algorithm is a frequently used alternative in shortest-path computations (Hart et al., 1968). Dijkstra's and A* shortest-path algorithms are both categorized amongst the one-to-all shortest-path algorithms. Many other algorithms besides the classic Dijkstra and A* algorithms have been developed, for example, the one proposed by Geisberger et al. (2008) who make use of contraction hierarchies which makes computation much faster.

The "*simplest path algorithm*" proposed by Duckham and Kulik (2003) is one of the alternative heuristics to the shortest-path algorithm discussed above. The authors argue that complexity of route instructions may be as important as total costs used to travel paths. Therefore, they develop an algorithm that offers the advantages of easier description and execution of the simplest path associated with a marginally longer path length compared to the shortest path.

2.4 Assessment of trajectory similarity

The assessment of similarity between different trajectories is an important tool that is, inter alia, used by researchers for method validation purposes and consists of comparing a modeled path to an observed path (Lou et al., 2009; Newson and Krumm, 2009; Pelekis et al., 2011), or, for data mining approaches (Giannotti et al., 2007; Laube et al., 2011). In particular, the research domains of computational geometry (Alt, 2009; Buchin et al., 2011) and time series analysis have been extensively involved in the development of concepts to analyze trajectory similarity (Dodge et al., 2012; Vlachos et al., 2004). To assess the resemblance between two objects, a frequently used approach is to assess the actual dissimilarity by quantifying the distance between them (Faloutsos et al., 1997). In similarity analysis, the two concepts of whole matching and subsequence matching are distinguished (Agrawal et al., 1993). In the former case complete trajectories and in the latter case subsets of trajectories are compared.

A set of trajectory similarity measures is presented in the following two sections. The focus is on spatial measures in Section 2.4.1, which assess the similarity of trajectories purely based on their geometric shapes in space, and on spatio-temporal measures in Section 2.4.2, which assess the similarity considering both the spatial and temporal dimensions of the objects (Dodge et al., 2012).

2.4.1 Spatial similarity measures

Many spatial similarity measures are based on the Euclidean distance. A very prominent measure, the Hausdorff distance, assesses the similarity between two sets of points A and B that represent, for example, curves (Alt, 2009). The directed Hausdorff distance $\vec{D}_H(A,B)$

is for all points a of A , the maximum of their minimum distance to B . The minimum distance of a point $a \in A$ to B is the shortest distance between a and all points of B . $\vec{D}_H(A,B)$ is expressed in the following Equation 1, whereas $\| \cdot \|$ represents the distance metric used, in this case the Euclidean distance:

$$\vec{D}_H(A,B) = \max \{ \min \{ \|a - b\| \mid b \in B \} \mid a \in A \} \quad (\text{Equation 1})$$

The bidirectional Hausdorff distance $D_H(A,B)$ is a symmetric measure and maximizes the directed Hausdorff distance from A to B and vice versa (Alt, 2009). It is expressed by the following Equation 2:

$$D_H(A,B) = \max (\vec{D}_H(A,B), \vec{D}_H(B,A)) \quad (\text{Equation 2})$$

The average Hausdorff distance such as used in Guerra and Pascucci (2005) computes the average minimum Euclidean distance between the points of set A to set B , or vice versa.

The similarity of network-bound trajectories is frequently assessed in terms of edge alignment. Newson and Krumm (2009) propose a measure which assesses the similarity between a modeled (matched) and a ground truth trajectory (correct route, see Figure 1). It is computed by dividing the sum of the total length of erroneously subtracted (d_-) and the total length of erroneously added (d_+) edges of the matched route (in comparison to the correct route) by the length of the correct route (d_0). Lou et al. (2009), who use a comparable definition of similarity as Newson and Krumm (2009), compute the number (length) of correctly identified road segments divided by the total number (length, respectively) of identified road segments in order to assess the quality of their map-matching algorithm.

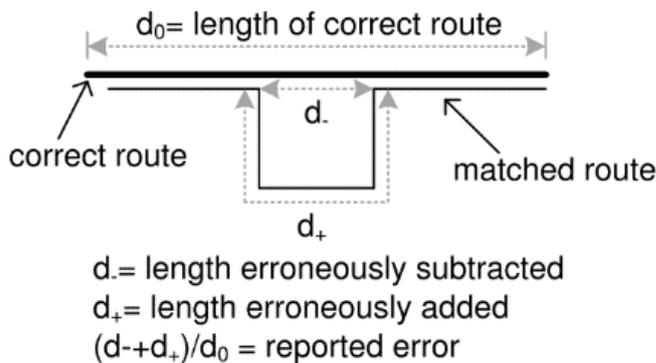


Figure 1: Schematic representation of similarity measure used by Newson and Krumm (2009)

Pelekis et al. (2011) propose an aerial similarity measure designated as locality in-between polylines (LIP). In Figure 2, the areas between two consecutive intersections resulting from the overlay of two trajectories are computed. Subsequently, the areas weighted according to the length of the respective sub-trajectories are summed, giving an indication of the distance between two trajectories.

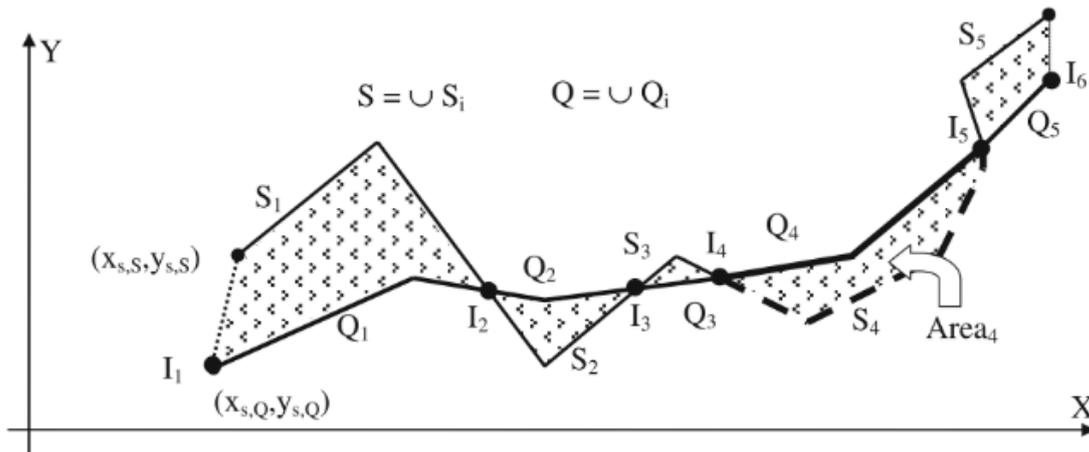


Figure 2: Locality in-between polylines (Pelekis et al., 2011)

An edit distance-based method such as the one used in Yin and Wolfson (2004) assesses the smallest number of insertions, deletions and substitutions (edit operations) required to change a trajectory to the one to which it is compared. Many other approaches, e.g., dynamic time wrapping or longest common subsequences as described in Vlachos et al. (2004, 2002), are used to assess the spatial similarity between trajectories.

2.4.2 Spatio-temporal similarity measures

The assessment of spatio-temporal similarity is complex and little research has focused on spatio-temporal similarity measures so far (Dodge et al., 2012). According to Dodge et al. (2012), most of the spatio-temporal similarity approaches employ a notion of Euclidean distance. A straightforward approach is to assess the dissimilarity between two time-referenced trajectories by measuring the average Euclidean distance between the points composing each trajectory at corresponding times (Buchin et al., 2011; Nanni and Pedreschi, 2006).

Fréchet distance is among the most frequently used similarity measures that consider the order of the points composing a trajectory besides its geometry (Alt and Godau, 1995; Alt, 2009; Brakatsoulas et al., 2005; Buchin et al., 2010, 2009). Fréchet distance is a similarity measure from computational geometry (Buchin et al., 2010). A common metaphor used to describe the Fréchet distance is a man walking his dog, both walking on their respective paths (Buchin et al., 2010; Dodge, 2011). While they are both allowed to control their speed, they are not allowed to move backward. The Fréchet distance for the two paths is defined as the minimum length of a leash that is necessary along the walk.

Giannotti et al. (2007) present a trajectory mining approach that aims at identifying similar trajectories in large GPS or GSM trajectory databases. Thereby, identical trajectory segments are identified that have visited similar places in the same order and used the same amount of time for the travel in between.

2.5 Identification of research gaps and research questions

As set out in Section 2.3, CDR data have already been extensively studied in many different contexts, but with the focus mainly on spatial patterns of human activities (e.g., defining home / work locations or tourism destinations), and less on the actual movement expressed by individuals (e.g., Ahas et al., 2010b, 2008b). The existing studies that try to assess mobile phone users' mobility on the basis of CDR data often use proxies (e.g., radius of gyration, in order to describe a typical range of a mobile phone user) (Frias-Martinez et al., 2010; Järv et al., 2014; Wang et al., 2011; Yuan et al., 2012). These proxies are used, inter alia, to characterize people according to their area of influence or to investigate whether there are correlations, e.g., between mobile phone usage or social network and the mobility of a person. A few researchers (Blumenstock, 2012; Csáji et al., 2013; González and Barabási, 2007), who reconstruct trajectories, connect the locations of consecutive phone activities with straight lines and do not make use of a road network, despite the evidence that most human movement is network-bound (Brinkhoff, 2002). The methods developed that make use of a road network in their research, work only in very specialized (constrained) settings (e.g., between a specific origin and destination) (Bar-Gera, 2007; Doyle et al., 2011; Järv et al., 2012).

The study of movement is of special interest to understand activities and processes occurring in the geographic space as well as for applications oriented towards traffic supervision and management. Many mobility studies, however, in ITS rely on GPS data or active mobile positioning data typically featuring a higher temporal resolution and / or mobile positioning data that have more precise location information, such as signal strength or RSSI (Asakura and Hato, 2004; Promnoi et al., 2009; Waadt et al., 2009; Zuo et al., 2012). These kinds of data do not share the typical features of CDR data that are low in temporal and spatial resolution and therefore their methods are not directly applicable to the CDR data.

The main advantages of CDR data are that they are available for a large fraction of the population, over a long time period and do not involve any active human tracking with additional devices (Ahas et al., 2010a; Caceres et al., 2007; Montoliu and Gatica-Perez, 2010). If the interest is in the movement of people, the CDR data need to be processed and a few assumptions (e.g., movements are bound to a network, use of shortest path to travel between locations) need to be made in order to be able to create trajectories from the low spatial and temporal resolution data (Kracht, 2004; Ratti et al., 2006). If the trajectories can be reasonably reconstructed, there is considerable potential in CDR data to assess overall mobility of a country and not only the mobility of a test sample such as described in Brakatsoulas et al. (2005). So far, relatively little methodological work has been published on how to reconstruct trajectories from CDR data.

One of the aims of this thesis is to develop techniques to reconstruct individuals' movement behavior in geographic space on the basis of a test CDR data set of 6 mobile phone users in Estonia over a one month period. Furthermore, the methods developed in the study are validated by assessing the accuracy of the reconstructed trajectories by comparing them to GPS data of the same users having a much finer temporal and spatial granularity. This thesis is therefore a response to the urgent need for validation of CDR methods as ascertained by Smoreda et al. (2013). The last aim of this thesis is to investigate which properties

of the CDR data have an impact on the accuracy of the reconstructed trajectories. The intention is to make statements regarding CDR data conditions under which a better quality of trajectory reconstruction is to be expected. To address the above-mentioned aims the following research questions (RQs) will be examined:

RQ 1: How can mobile phone users' trajectories be reconstructed from sparsely sampled CDR data?

RQ 2: In order to validate the trajectory reconstruction methods developed in this study, what level of similarity can be achieved by comparison of the reconstructed trajectories with higher resolution GPS trajectories of the same journeys?

RQ 3: Which properties of the CDR data, such as sampling properties or trajectory length, affect the accuracy of the reconstructed trajectories?

3 Overview of the overall workflow and the software used

This chapter gives a brief overview of the Chapters 4 - 6, which constitute the methodological part of this thesis, and introduces the software that was used.

Chapter 4: Description and pre-processing of the data

A description of the main characteristics of the initial CDR, GPS and OpenStreetMap (OSM) road network is given in the first part of Chapter 4. In a second part, the pre-processing and filtering steps applied to the different data sources are described. The positioning data are partitioned into temporal units suitable for analysis. The road network needs to be made routable and implemented as a graph, in order to be able to use it for routing purposes.

Chapter 5: Development of methods to reconstruct trajectories from CDR data

In Chapter 5, methods are developed to reconstruct paths from sparse CDR data. In a first step, CDR fixes are matched to nodes on the graph, and in a second step, shortest paths are computed between the identified nodes.

Sections 6.1-6.4.1: Validation of trajectory reconstruction methods by comparison of reconstructed paths to ground truth

The validation of the developed trajectory reconstruction methods is done by comparing the reconstructed paths to the ground truth which is derived from the GPS data. Therefore, a range of similarity measures is proposed and implemented.

Section 6.4.2: Impact of CDR data properties on accuracy of trajectory reconstruction

In Section 6.4, the similarity measures are investigated with respect to the properties of the CDR data that the trajectory reconstruction is based on. The aim is to make statements regarding data conditions (e.g., a notion of minimum number of CDR fixes) under which a higher accuracy of trajectory reconstruction is to be expected.

Software used

For the implementation of the above-mentioned steps, the Integrated Development Environment (IDE) Eclipse (2014) based on Java (2014) was used. For many spatial functionalities and operations GeoTools (2014) – an Open Source Java Library – was accessed. Furthermore, Maven (2014), a Software Project Management Tool, was used. For visualization purposes, besides mapping options provided by GeoTools, ArcGIS 10.2 (2014), a commercial GIS software of ESRI (2014), was used. Microsoft Office Excel 2013 (Microsoft, 2014) and the IDE RStudio (2014), based on the R environment (R-project, 2014), were applied for statistical analyses and visualizations.

4 Data and pre-processing

4.1 Overview

The following chapter describes the initial positioning and road network data and the pre-processing steps applied to the different data sources.

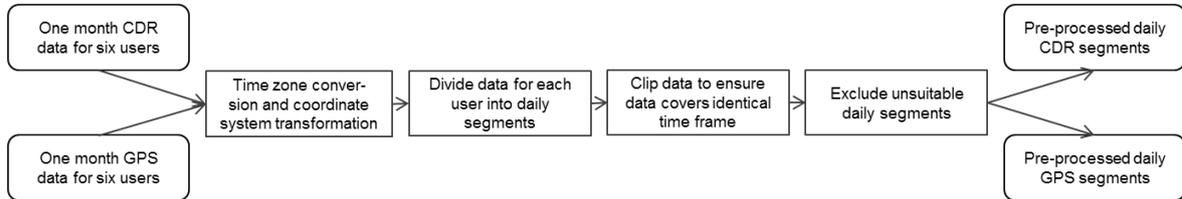


Figure 3: Workflow of the pre-processing of the positioning data

The original one month CDR and GPS data of the six mobile phone users are described in Section 4.2. The following pre-processing steps, as visualized in the workflow in Figure 3, are undertaken for the positioning data: The time zone is converted to make it identical for the GPS and the CDR data and the coordinate system is transformed from a spherical to a Cartesian one (Section 4.3.1). The one month data for the six users is subsequently divided into analyzable daily segments (Section 4.3.2). The daily CDR and GPS files are subsequently clipped according to the time frame of each other, in order to ensure that the analyzed time periods of the two data sources are identical and therefore comparable (Section 4.3.3). In the last step, the daily segments, unsuitable for the analysis, are excluded from the data sample (Section 4.3.4).

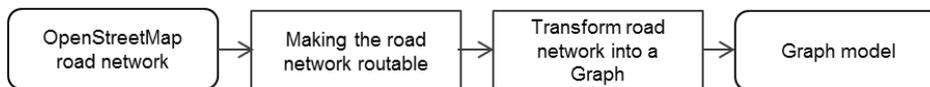


Figure 4: Workflow of pre-processing of the OSM road network

Figure 4 shows the workflow of the OpenStreetMap (OSM) data pre-processing, which is described in Section 4.4. In a first step, the original road network is made routable (Section 4.5.1). In a second step, the routable road network data is transformed into a graph model (Section 4.5.2).

4.2 Positioning data

The CDR and GPS data of six mobile phone users from Estonia – referred to as users 1, 3, 4, 5, 6, 7² – for the month of June 2013 have been used for this study. Every time a mobile phone user initiates a phone activity (such as a call, a SMS, or a MMS)³ the location of the antenna that routes the phone activity and the starting time of the respective interaction are registered as one entry into a log file. A GPS device, which is integrated into a user’s mobile phone, continuously registers the locations and the corresponding times. In the following,

² The users expressed their consent to use the data. The data of user 2 were not used for this study because it originates from a different mobile phone operator.

³ In the following, the two terms “CDR fixes” and “phone activities” are used interchangeably both referring to initiated – not received – calls, SMS and MMS by a user.

the aforementioned registered CDR and GPS entries are referred to as CDR and GPS fixes, respectively. Table 1 shows the number of CDR and GPS fixes available of the different users. It can be seen that there are considerable differences in terms of number of CDR fixes and GPS fixes. While user 1 has 75 CDR fixes in total, which corresponds to 2.5 fixes per day on average, user 4 has 292 fixes in total, which corresponds to 9.7 fixes per day on average. It is also observable that the number of CDR fixes does not correlate with the number of GPS fixes. User 1, for instance, with the highest number of GPS fixes has the lowest number of CDR fixes.

Table 1: Initial number of CDR and GPS fixes per user

User	Number of CDR fixes	Number of GPS fixes
1	75	207'889
3	219	139'208
4	292	184'617
5	79	135'519
6	152	110'382
7	169	26'262

4.2.1 CDR data

Table 2 shows an excerpt of the original CDR data. Notably, the CDR data features (a) the user id, (b) the date and time (Eastern European Summer Time (EEST) with the precision of a second) of the outgoing phone activity, and (c) the longitude and (d) the latitude in World Geodetic System 1984 (WGS 84) coordinate system of the antenna to which the mobile phone was connected during the activity. Good descriptions of the CDR data, similar to the one used in this study, are provided in Ahas et al. (2010b, 2009) or Järv et al. (2014).

Table 2: Excerpt from the original CDR data of user 6

Non-identifiable user ID	Start Time (EEST)	Longitude	Latitude
6	01.06.2013 11:49:15	26.7305472	58.3802722
6	01.06.2013 13:40:18	26.7305472	58.3802722
6	01.06.2013 17:03:35	26.73	58.3669444
6	01.06.2013 18:54:02	26.72025	58.3802444
6	01.06.2013 22:31:06	26.7205555	58.3813888
6	01.06.2013 22:53:48	26.7202666	58.3925

User ID

CDR data are pseudonymous and with the consent of a mobile phone user it is possible to identify his unique pseudonym. The CDR and GPS data owners are matched when already known GPS users ask their billing extraction from the mobile operator and then this is used to find the user ID from a large pseudonymous CDR database.

Temporal Resolution

The temporal resolution of the CDR data is highly dependent on the frequency of phone activities a mobile phone user initiates (but not receives in the case of this study). There might be a gap of a couple of seconds up to several hours between two consecutive phone activities. On average, the users in the data sample make about 6 phone activities per day. Therefore the temporal resolution of the data, on an individual level at least, is rather coarse.

Spatial Resolution

The spatial accuracy of the CDR data is inherent to the structure of the antenna network, which reflects the population density patterns and the transportation infrastructure (Ahas et al., 2010b). Voronoi cells⁴ define the area closest to each antenna. They are a frequently used approximation of the area where the mobile phone user can be assumed to be located during the phone activity (Ahas et al. 2009). If the antenna is very crowded or the visibility of the antenna is disturbed, the mobile phone might switch to any other antenna in the neighborhood (Ahas et al., 2010b). The spatial accuracy is dependent on the sizes of the Voronoi cells. These again are dependent on the density of the antennas. According to Ahas et al. (2008b) in Estonia's biggest cities Tallinn, Tartu and Pärnu, location accuracies between 100 and 1000 m may be expected. In suburban regions, spatial accuracies vary between 450 m and 2 km. Rural areas, which are mostly unpopulated such as Estonia's remote wetland areas, have a spatial accuracy between 1.5 and 20 km (Ahas and Laineste, 2006). Based on the calculation of the theoretical positioning error on 180'000 positioning measurements, Ahas et al. (2007b) show that the accuracy of mobile positioning data in Estonia is within 1000 m for 61% of the positioning points in urban areas and within 3000 m for 53% of the positioning points in rural areas.

Availability of the data

CDR data is automatically stored for all mobile phone users in any mobile phone operator's network according to the billing purposes and data retention directives from the EU (e.g., European Parliament, 2002). TNS EMOR (2014) omnibus survey in Estonia considering mobile phone penetration initiated by Ahas et al. (2010b) showed that 95% of the population has a mobile phone. Private company Positium LBS (2014) purchases this billing information in a pseudonymous form and this information is delivered with a fixed interval to Positium's servers. Currently Positium LBS handles mobile positioning data from all three Estonian mobile phone operators: EMT, Elisa and Tele2. The one month CDR data for the six users in this study is provided by Positium LBS in collaboration with its long-standing partner in academia, the Geography Department of the University of Tartu.

Summary

A summary of the most important features of the CDR data is given in Table 3.

⁴ Given a set of discrete points in a metric space, the Voronoi cell of such a point is the set of points in that space that are closer to that point than to any other point, a Voronoi diagram is then the set of all those Voronoi cells (Aurenhammer, 1991).

Table 3: Summary of the most important features of the CDR data

Spatial Resolution	Temporal Resolution	Availability
<ul style="list-style-type: none"> – Area of Voronoi cell (approximation) – Dependent on density of mobile phone network: – Urban area: Ø 1000m – Rural area: Ø 3000m 	<ul style="list-style-type: none"> – Dependent on number of phone activities – 6.17 phone activities / day for original data in this study 	<ul style="list-style-type: none"> – Potentially of total population (95% mobile phone penetration in Estonia) – 6 test users in this study

4.2.2 GPS data

For each user a file of one month GPS data collected with an Android application for the month of June 2013 is available for this analysis. Table 4 shows an excerpt of the original GPS data of user 1. A GPS entry features the following attributes: (a) time of observation in milliseconds elapsed since start of Unix epoch (00:00:00 Coordinated Universal Time (UTC) on January 1, 1970) with the precision of a second, (b) the longitude, and (c) the latitude of the mobile phone at the respective time, as well as (d) the speed of the mobile phone in m/s. The speed measurements are not used in this study. The other attributes are further described in the following.

Table 4: Excerpt from the original GPS data of user 1

Time (Unix timestamp)	Longitude	Latitude	Speed (m/s)
1371825786000	59.423379	24.7956574	2.125
1371825787000	59.4233822	24.7956957	1.6875
1371825788000	59.4233854	24.7957276	1.375
1371825789000	59.4233869	24.7957496	0.875
1371825790000	59.4233903	24.7957664	0.4375
1371825791000	59.4233928	24.7957846	0

Temporal Resolution

The temporal resolution of the GPS data is dependent on how much the person moves. If the mobile phone is stationary over a longer period, GPS positions are not further registered to save the battery life of the device. If the mobile phone is moved, the frequency of time-referenced points registered differs depending on the movement speed. The higher the speed, the more entries are registered. The GPS data in this study have an average time between consecutive GPS fixes of 19 s. The deviation, however, is very high.

Spatial Resolution

Reviewing the literature, horizontal positioning errors of GPS devices can be found between approximately 2 and 25 m, depending on (amongst other factors) the measurement conditions (environment) and the devices used (Haklay and Weber, 2008; Kuter and Kuter, 2010; Sigrist et al., 1999; Wing and Eklund, 2007). A typical horizontal positional error, which is indicated in a manufacturer's technical specifications, would be less than 7 m in 95% of the cases (Thales, 2005). It can be assumed that the horizontal positioning error in this study is approximately within the same range, without knowing the exact specifications.

Availability

GPS data is only gathered if the respective people are in accordance with actively being traced and have signed an agreement. Mobility Lab of University of Tartu is collecting GPS data of approximately 60 volunteers in Estonia (MobilityLab, 2014). The number of participating people is increasing as new volunteers are continuously found. The mobile phones of the respective people are provided with a GPS chip and an Android platform, where special software is installed. This software collects and transmits the data securely into the Mobility Lab database when an internet connection is available.

Summary

A summary of the most important features of the GPS data is provided in Table 5.

Table 5: Summary of the most important features of the GPS data

Spatial Resolution	Temporal Resolution	Availability
– Approx. < 7 m	– Depending on speed – Approximately every 19 s	– People need to be actively traced – The same 6 test users

4.3 Pre-processing of positioning data

4.3.1 Time zone conversions and coordinate system transformation

To be able to compare CDR and GPS data, the time indications need to be within the same time zone. The CDR data which is in Eastern European Summer Time and the GPS data which is in UTC are therefore transformed to the computer time, which is Central European Summer Time (CEST). Since the data was collected in June 2013, summer time is relevant. Therefore one hour is subtracted from the time of the CDR data and two hours are added to the time of the GPS data. Due to the time transformation, the first CDR fix of user 3 – at 00:28:41 in EEST and at 23:28:41 in CEST – is preponed to the May 31, 2013. This CDR fix is manually deleted from the original CDR data file.

The original CDR and GPS datasets are converted from the geographic WGS 84 to the projected coordinate system Universal Transverse Mercator (UTM) zone 35N⁵ in which Estonia is mostly contained. UTM uses a 2-dimensional Cartesian coordinate system.⁶ Since the Cartesian coordinate system is built upon two perpendicular axes and the unit is m, distance calculations are facilitated.

4.3.2 Segmenting the CDR data

To calculate trajectories, the temporal frame of reference to investigate needs to be defined. Figure 5 shows a typical phone activity frequency distribution over the period of a week. A diurnal pattern is clearly distinguishable (Csáji et al., 2013; Järv et al., 2012). The threshold time that divides the data could be set to the time with the minimum mobile phone activity,

⁵ <http://spatialreference.org/ref/epsg/wgs-84-utm-zone-35n/> (accessed 10.5.2014)

⁶ For further information check the ArcGIS Resource Center http://help.arcgis.com/en/arcgisdesk-top/10.0/help/index.html#/Universal_Transverse_Mercator/003r00000049000000/ (accessed 10.5.2014)

which could be referred to as “functional midnight”. According to Figure 5, this would be between 4 and 6 o’clock, depending on whether it is a workday or a weekend day. For simplicity, in this study the CDR data is divided into daily segments based on the date. Consequently, the basic temporal unit that is investigated corresponds to one day, which starts at 00:00 and ends at 23:59. Given that the algorithm uses computer time the threshold is set to 24:00 CEST, which corresponds to 01:00 EEST. The time lag of one hour is unproblematic though because, as mentioned before, the functional midnight would be later than midnight.

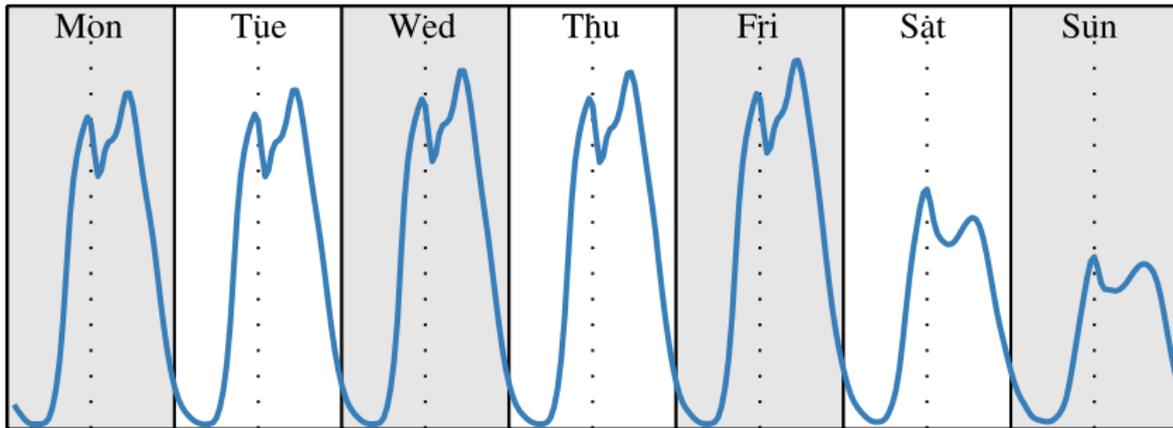


Figure 5: Average phone activity dynamics (Csáji et al., 2013)

The fixes belonging to one unit of analysis are stored in daily segments maintaining all spatial and temporal attributes, as well as the mobile phone user’s ID. The files are designated in the following manner: x_{1306dd} (“x” representing the user number, “13” the year of 2013, “06” the month of June, and “dd” the day of the month). Through the partitioning, 156 and 165 separate CDR and GPS daily segments are created, respectively. Table 6 shows the statistics for the daily CDR files of each user. The average number of fixes per day varies between 3.2 and 9.7 for the different users. Comparing the average number of fixes per day to the average number of fixes from different locations (via different antennas) per day, it is observable that the former is considerably higher than the latter. This second number is important, because consecutive fixes with identical position are of no use for the trajectory reconstruction methods used later on.

Table 6: Statistics for daily CDR segments per user

User	Number of days	Average number of fixes per day	Average number of fixes from different locations per day	Average total time between first and last fix [min]	Average mean time between consecutive fixes [min]	Average total distance between first and last fix [m]	Average mean time between consecutive fixes [m]
Only daily segments with $n > 1$ considered for these statistical values							
1	21	3.6	2.4	420.6	169.9	54'167.6	10'617.2
3	24	9.1	3.0	584.2	79.1	37'341.7	5'994.7
4	30	9.7	4.9	624.8	93.0	90'744.9	11'532.2
5	25	3.2	1.8	313.7	124.3	27'116.9	4'795.5
6	27	5.6	3.7	540.3	118.8	37'483.9	6'820.5
7	29	5.8	3.2	456.2	126.0	4'810.0	1'093.7

4.3.3 Clipping the CDR and GPS daily segments for similar time frames

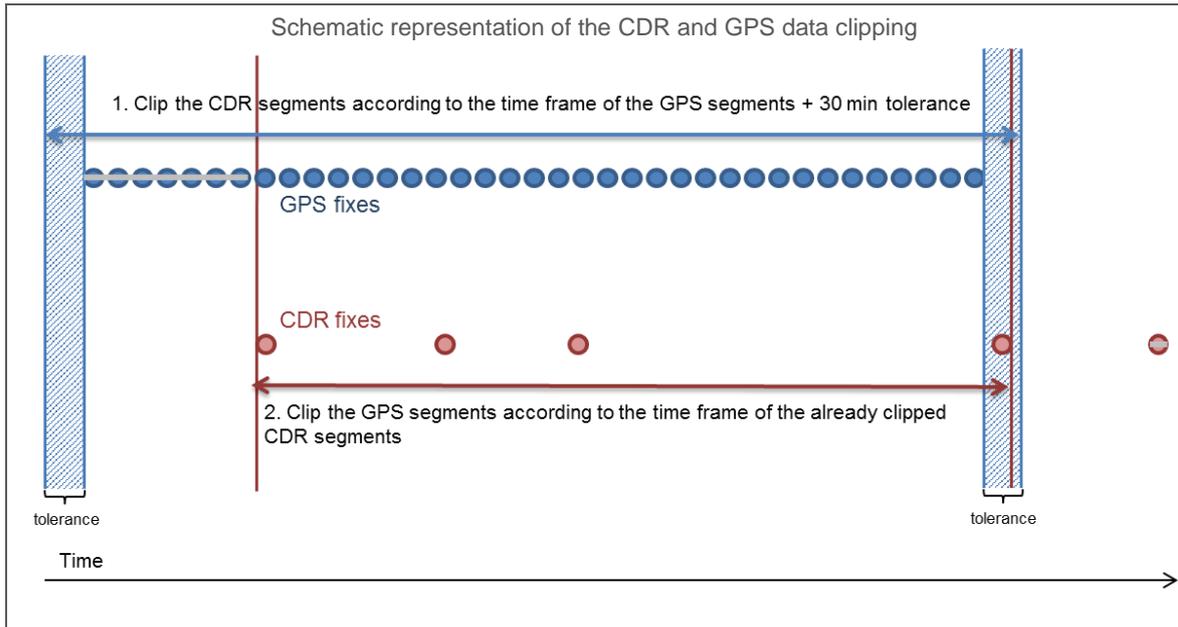


Figure 6: Principle of the clipping of the daily segments: fixes crossed out in grey are discarded due to lack of corresponding data of the opposite data source

The daily segments, generated as described in the previous Section 4.3.2, are clipped according to the schematic representation in Figure 6. In a first step, the CDR data are clipped to the time frame given by the first and the last fix of the GPS data plus a 30 min tolerance. The tolerance guarantees that CDR fixes shortly before the first GPS position and after the last GPS position are not excluded. Without such a tolerance, the fourth CDR fix, lying in the blue hatched area of the schematic representation in Figure 6, would have been discarded, although this fix would have been informative for the prediction of the movement registered by the GPS device. In a second step, the GPS fixes are clipped to the time frame given by the first and the last fix of the already clipped CDR data.

The reason for the clipping is that validation by comparison of the two data sources is only logical for corresponding time frames of the GPS and the CDR data. It does not make sense to keep CDR data for time periods with no ground truth (GPS) data available. The opposite is also true, no ground truth data should be kept for time periods when no path reconstruction is feasible due to lack of CDR data. As consequence of the clipping, it is possible that complete CDR daily segments are deleted because there is no corresponding GPS data of the respective time period available and vice versa. Out of 156 CDR daily segments, 17 are excluded because there was either no corresponding GPS segment or no temporal overlap of the CDR and the GPS daily segments. Hence, 139 GPS – CDR comparison cases are remaining after the clipping.

4.3.4 Excluding daily segments unsuitable for reconstruction

The disqualification criteria for the daily segments vary by the two data types that are used. These are described in the following two sections. Obviously, if a CDR daily segment is excluded for further analysis, the corresponding GPS segment is excluded as well and vice versa.

4.3.4.1 Exclusion of unsuitable daily CDR segments

As an absolute precondition for trajectory reconstruction, at least two fixes from different places per day are required. Therefore, daily segments consisting of less than two fixes from different places are excluded from the data sample. In this way, 37 daily segments are disqualified, 20 of which consisted of only one CDR fix. Hence, out of the 139 comparison cases, 102 are left for further investigation.

4.3.4.2 Exclusion of unsuitable daily GPS segments

To exclude daily segments in which the original GPS points show excessively large spatial gaps between consecutive points, a maximum tolerable gap length is defined. To define such a threshold, all the gap lengths between consecutive fixes of all the clipped daily GPS segments, which are supposed to serve as ground truth, are extracted. Subsequently, the extracted gap lengths are classified into groups of 50 m each. Figure 7 shows a frequency diagram for the different gap length classes in a logarithmic scale. The orange strokes indicate values where changes in the frequency distribution are discernible. They represent candidate thresholds. In this study, the maximum gap length tolerated is set to 700 m, since this is the stricter threshold between the two candidates. If this threshold is exceeded, the respective daily segment is excluded for further analysis.

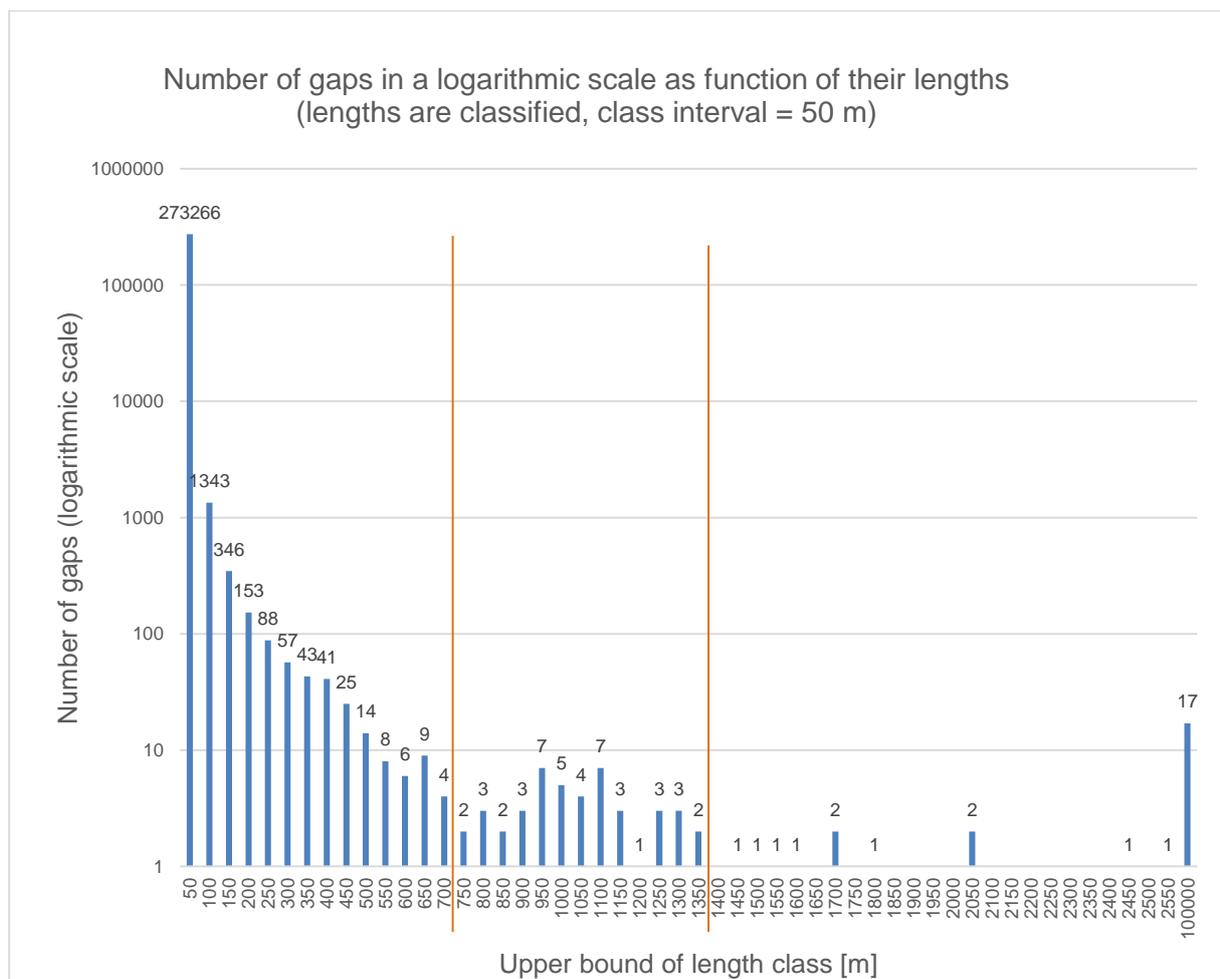


Figure 7: Frequency diagram representing the number of gaps as a function of their lengths

Figure 8 shows the number of affected daily segments as a function of the maximum tolerated gap length between consecutive GPS fixes. With a threshold set to 700 m, 21 segments are excluded (as indicated with the orange stroke) from the 102 daily segments. Hence, the data sample is reduced to 81.

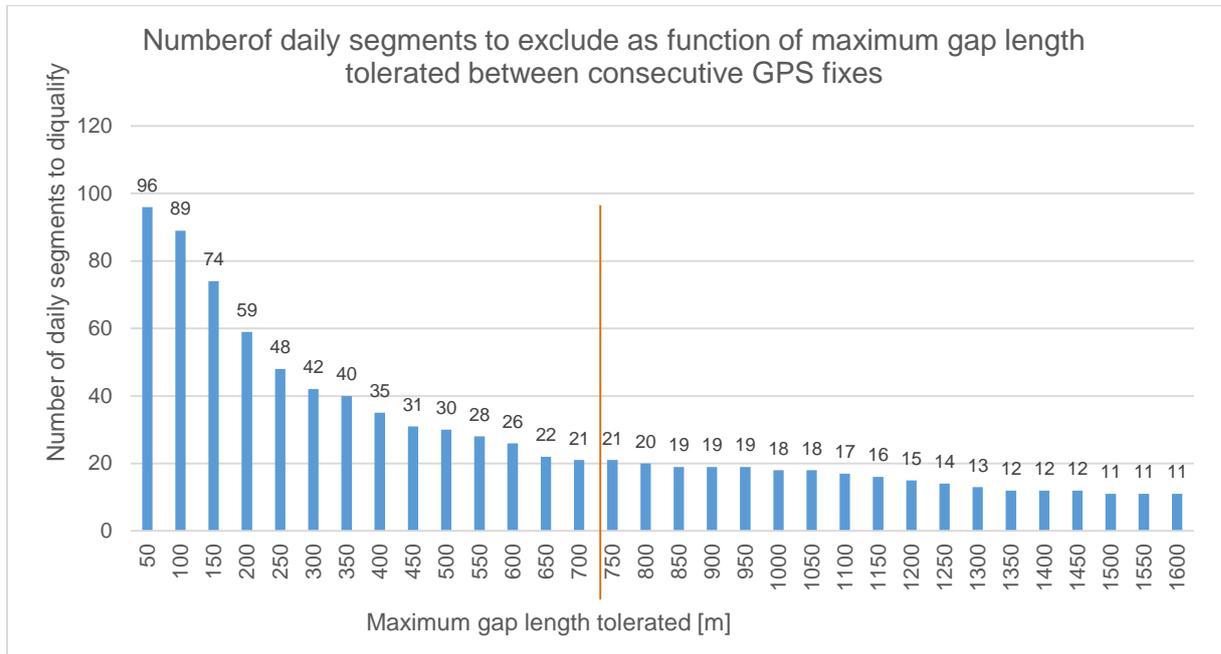


Figure 8: Diagram representing number of daily segments to exclude as a function of the tolerated maximum gap length between consecutive GPS fixes

4.4 OSM road network data

The road network data is from OpenStreetMap (OSM, 2014). OSM is the “Wikipedia of maps”. It is a community project and its content is generated by volunteering contributors (Haklay and Weber, 2008). The biggest advantage of OSM is that the data are freely available to everyone. One of the disadvantages is that there is no guarantee of comprehensiveness, consistency and correctness (Mondzech and Sester, 2011). Haklay (2010) reported a geometric accuracy of the OSM data between 5 and 20 m for the city of London. The coverage of OSM data differs highly according to the region; especially in rural areas there are still blank parts.

The features in the OSM data are tagged with so-called key-value combinations. A key describes the broad class of the feature and the value gives further details about the feature that was classified with a tag (OSM, 2014). The key “highway”, for example, describes any road connecting one location to another that has been paved or otherwise improved to allow passage by motorized vehicles, cyclists, pedestrians and others (OSM, 2014). The value, which might be “motorway”, “primary”, “residential”, etc., gives further details about the type of the feature generally classified as “highway”.

OSM data for most of the countries are downloadable as shapefiles on Geofabrik (2014). Geofabrik provides the OSM data in a zip file per country or a bigger region in the osm.bz2 data format which yields an OSM XML after decompression. The Estonia “estonia-latest.osm.bz2” zip file for this thesis was downloaded on October 30, 2013.

4.5 Pre-processing of OSM road network data

4.5.1 Making OSM road network routable

On the Geofabrik homepage it is also possible to directly access the roads of Estonia as ESRI compatible shape file. This file, however, is not routable, which implies that it is not suitable for graph processing, such as a shortest-path computation. Since a shortest-path algorithm needs to be implemented, the original OSM data needs to be accordingly processed. The program OSM to RouteWare v.1.07c downloaded from Routeware (2014) was therefore used. According to RouteWare (2014) the program is designed to make the OSM data routable and highlight errors in the data. Graser and Straub (2013) give an overview of the characteristics a road network should feature, in order to be suitable for routing purposes. The unzipped “.osm” file containing all data from Estonia serves as input to the OSM to RouteWare program. As output a shapefile called “roads.shp” is generated. Table 7 gives an overview of the attributes that come with each road feature represented as polyline.

Table 7: Attributes of routable road features

Attribute	Description
OSM ID	Unique number of the OSM feature (e.g., “128743565”)
Name (optional)	Name of the feature (e.g., “Raudtee”)
Reference (optional)	Reference of the feature (e.g., “Interstate 12” or “A1”)
Type	Value of OSM key “highway” or ferry (e.g., “secondary”)
Attribute in RW Net format	A number defined by RouteWare describing the road class
Max Speed (optional)	Speed limitation in in km/h, available for the features tagged with the OSM key “maxspeed” (e.g., “90”)
Duration	Time to travel the edge in min, only available for a few ferry edges (if not available duration = 0)

When visually comparing the original roads shape file of OSM to the roads shape file made routable by OSM to RouteWare, the following main differences are discernible:

- Highways tagged with the following values are excluded from the original roads file: ‘path’, ‘track’, ‘pedestrian’, ‘cycleway’, ‘footway’, ‘proposed’, ‘construction’, ‘raceway’, ‘bridleway’. In summary, roads that are not traversable by cars are excluded.
- Lines tagged with the value ‘ferry’ out of the class with the key “route” are added to the routable road network. The original roads file from OSM only contains features with the key “highway”.
- The lines in the routable shape file are split at all intersections. This is needed for the routing. Ferry edges, however, are never split. This makes sense since ferries cannot be switched once one is upon them.

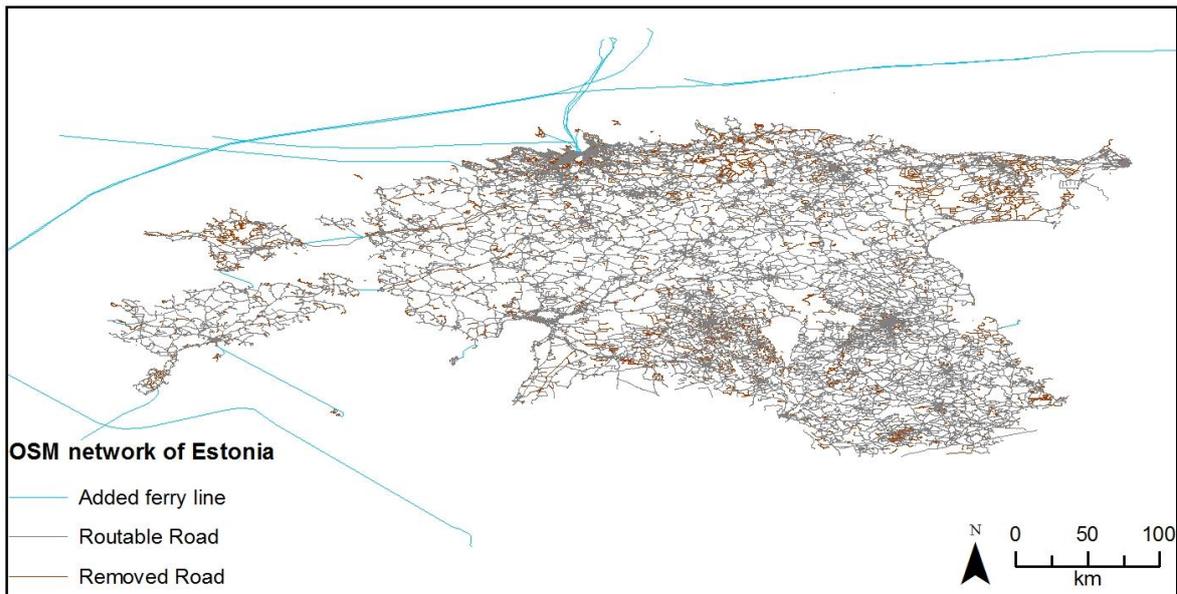


Figure 9: OSM⁷ road network of Estonia, showing the added and removed lines

Figure 9 shows the edges which are added (blue, ferry lines) or removed (brown, roads not traversable by cars). In terms of length, the original shape file with a total length of 48'700 km is 15% longer than the routable shape file with a total length of 41'200 km.

4.5.2 Transforming OSM road network to a GeoTools graph model

The polylines stored in a roads shape file, usable for routing purposes, must now be transformed into a graph model. A graph consists of a set of nodes connected by edges. The edges might have additional attributes such as maximum speed, driving directions, etc. Cao and Krumm (2009) present in their paper how a graph can be generated out of trips registered as GPS traces.

The *LineStringGraphGenerator* of GeoTools (2014) is used to generate an undirected graph – comparable to the graph representation used in Crucitti et al. (2008) – out of the shape file consisting of multiline features representing roads. A detailed description is available on the homepage⁸. Before turning the polylines into the edges of the graph, their coordinates are transformed from the global coordinate system WGS 84 to the Cartesian coordinate system UTM Zone 35N, which is also used for the positioning data (see Section 4.3.1). The intersections between the edges represent the nodes of the graph. The generated graph consists of 99'299 nodes and 123'976 edges. Since the graph is undirected, there is no information available about the direction of the traffic on a specific edge.

⁷ © OpenStreetMap contributors for all the maps in this thesis, <http://www.openstreetmap.org/copyright> (accessed 13.6.2014)

⁸ <http://docs.geotools.org/latest/userguide/extension/graph/index.html> (accessed 10.5.2014)

5 Trajectory reconstruction based on CDR data

5.1 Overview

The focus in this chapter is on the development of methods to reconstruct trajectories from the pre-processed CDR segments. The flow chart in Figure 11 illustrates the steps the trajectory reconstruction is composed of.

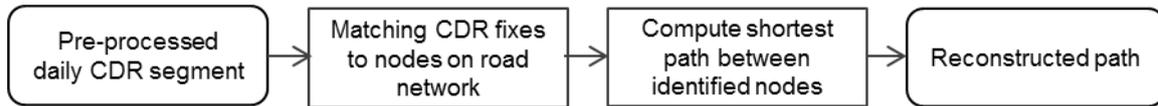


Figure 11: Workflow of trajectory reconstruction

Firstly, the pre-processed daily segments of CDR fixes are matched to a node on the road network. Seven different rules are applied in order to find the most probable node on the graph to which the user realistically might have been closest. Secondly, the shortest path is computed between the identified nodes on the network. This results in a set of edges that constitute the reconstructed path.

5.2 Matching CDR data to a node on the network

A straightforward but not very accurate approach, to reconstruct the trajectories, would be to connect the consecutive locations where mobile phone activities took place. There are at least two critical issues regarding this approach: Firstly, the location of the cell tower is, most of the time, not a good approximation for the place where the actual phone activity was initiated. Secondly, as Brinkhoff (2002) states, most human movements take place on a network. Thus, it makes sense to take some kind of network into consideration. The network is provided by the OSM road network data as described in Section 4.4.

In a first step, the CDR fixes need to be matched to the road network. Therefore, the following different map-matching (MM) approaches are proposed and further described in the following sections: A straightforward approach is to identify the node on the road network lying closest to the antenna position (Section 5.2.1). A reasonable assumption would be that a mobile phone user might be anywhere within the area that is closest to the respective antenna that routed the phone activity. This region is described by a Voronoi diagram (Section 5.2.2). Based on the Voronoi cells, it is possible to compute the center of gravity and subsequently find its closest graph node (Section 5.2.3). By intersecting the Voronoi diagram with the graph nodes, candidate nodes, representing the potential whereabouts of the mobile phone user, can be identified. Based on the characteristics of the nodes, the most probable location amongst the candidate nodes can be identified (Section 5.2.4). Finally, it is also possible to identify the candidate edges by intersecting the edges with the respective Voronoi cell, and subsequently, base the choice of the most probable edge on characteristics of the edges (Section 5.2.5).

5.2.1 MM method 1: Approach relying on proximity of nodes to antenna

The most straightforward approach to come from the antenna position to a node on the road network is to go through all the graph nodes and determine the node with the shortest Euclidean distance to the antenna (cf. point-to-point matching described in White et al. (2000)). The estimation of the location of the mobile phone in the place of the antenna is an inaccurate localization technique, but since it is a straightforward one, this technique is frequently used in literature (Csáji et al., 2013). The algorithm devised for MM method 1 works in the following manner:

Algorithm MM method 1: Find the closest graph node to an antenna

Input:

Graph graph

Antenna antenna //antenna which is supposed to be map matched to a graph node

Output:

Node nearestNode //nearest node of the graph to the input antenna

Initialize:

Double shortestDistance //initially set to a high number

Node nearestNode //initially set to zero

```

for (Node node : g.getNodes()) //iterate through all graph nodes
    Node currentNode = node;
    Double currentDistance = Euclidean Distance between currentNode and antenna
    If currentDistance < shortestDistance
        shortestDistance = currentDistance
        nearestNode = currentNode
  
```

Figure 12 shows the visualization of a subset of the data of user 4. The closest node, depicted in red, will be stored amongst the set of destination nodes used for the shortest-path computation in a further step.

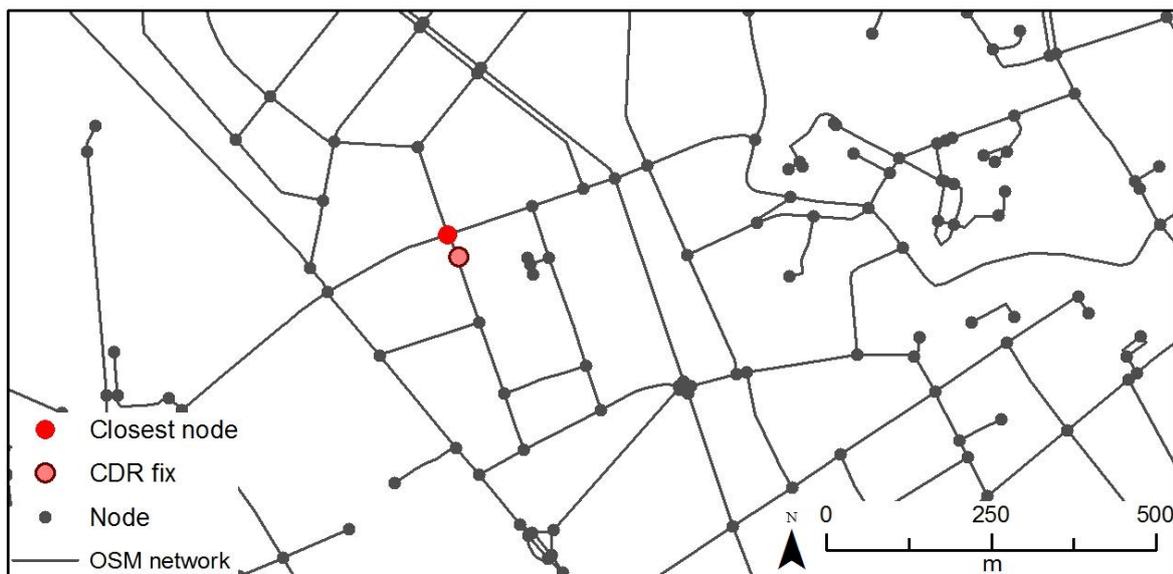


Figure 12: Node on the road network closest to the CDR fix (subset of CDR data of user 4, 03.06.2013)

5.2.2 Computing the Voronoi diagram

The map-matching methods described in the following Sections 5.2.3 - 5.2.5 are based on the identification of the Voronoi cells that surround each antenna. The Voronoi diagram is computed on the basis of the antenna locations extracted from the CDR files and additional antenna locations, which were provided, with a slight positional shift due to data security reasons, by Positium LBS (2014). The resulting static map of antennas consists of the 1052 antenna positions available in June 2013, of which 96 positions are extracted from the six one month CDR files. Figure 13 shows the Voronoi diagram computed with help of the *VoronoiDiagramBuilder*⁹ provided by GeoTools (2014). A Voronoi diagram tessellates a metric space into a cell for each antenna site based on the Euclidean distance (Aurenhammer, 1991). All points included in the resulting Voronoi cell surrounding a particular site are closer to this site than to all the other sites. Points lying on the segments of the Voronoi cells are equidistant to two antenna sites. A clear overview of how to construct a Voronoi diagram based on the antenna locations is given in Waadt et al. (2009).

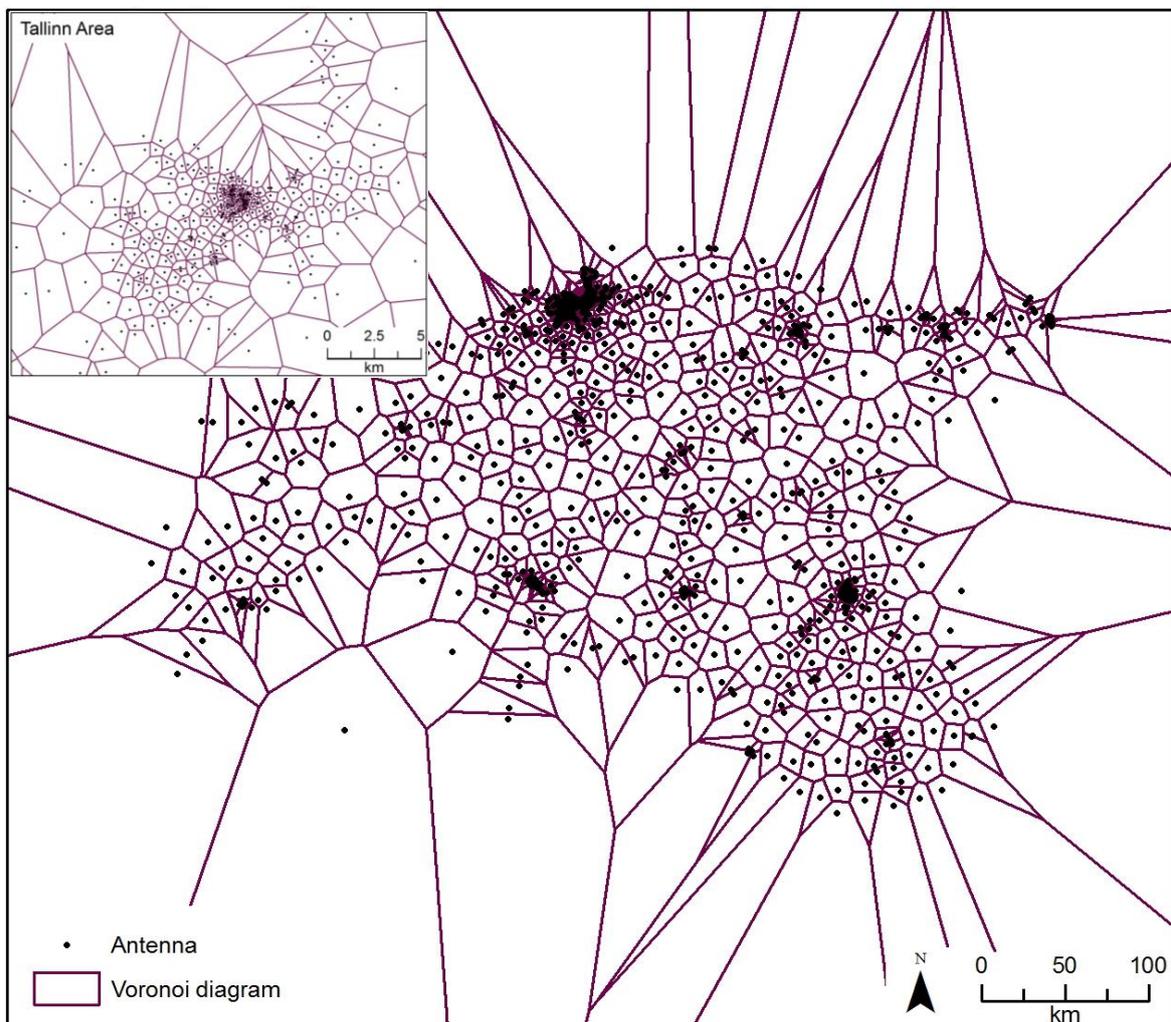


Figure 13: Voronoi cells describing the area closest to each antenna (Estonian mobile phone operator's antennas, partially slightly shifted, June 2013)

⁹ <http://tsusiatssoftware.net/jts/javadoc/com/vividsolutions/jts/triangulate/VoronoiDiagramBuilder.html> (accessed 11.5.2014)

Voronoi cells are a common approximation of the geographical areas of cell coverage (e.g., Doyle et al., 2011; Frias-Martinez et al., 2010; Waadt et al., 2009; Zang et al., 2010). They are easy to compute and result in non-overlapping regions, which facilitate the non-ambiguous assignment of an individual's location to the area of the Voronoi diagram. The approximation of a cell's geometry by a Voronoi diagram is based on the assumption of free room propagation, all antennas having equivalent radiated power, and both phone and antenna software will look for the best service (Waadt et al., 2009).

5.2.3 MM method 2: Approach relying on center of gravity of Voronoi cells

A common approach in literature (e.g., Doyle et al., 2011; Smoreda et al., 2013; Waadt et al., 2009) to approximate a user's location as a point is to use the center of gravity of a polygon. Similarly to Doyle et al. (2011), the centers of gravity of the Voronoi cells are used as an approximation for the location where the individual is assumed to be. Waadt et al. (2009) argue that the estimation error of the mobile station's location can be minimized by taking the cell's center of gravity instead of the location of the antenna.

The center of gravity was calculated using the *getCentroid*¹⁰ function provided by GeoTools (2014). Figure 14 shows the center of gravity in contrast to the antenna location. In a second step, the estimated location by the centroid function needs to be map matched to a node of the graph model. This is done in a fashion similar to that described in the algorithm in Section 5.2.1. The node lying closest to the center of gravity indicates the estimated location of the user and will be used as a basis for the path calculation in a further step.

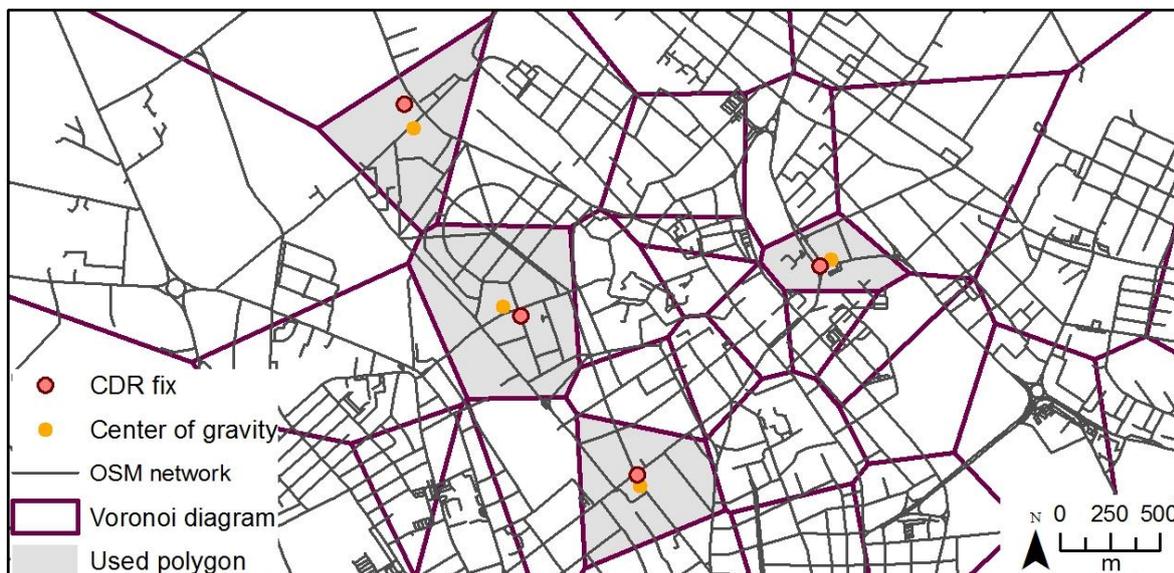


Figure 14: Center of gravity of the Voronoi cells used (subset of CDR data of user 4, 03.06.2013)

¹⁰ <http://www.vividsolutions.com/jts/javadoc/com/vividsolutions/jts/geom/Geometry.html#getCentroid%28%29> (accessed 12.5.2014)

5.2.4 MM method 3 and 4: Approaches relying on degree centrality of nodes

The two approaches described in this section are both based on a pre-selection of candidate nodes. Therefore candidate nodes that are contained by a Voronoi cell are identified and stored in sets of candidate nodes. One cell contains on average 94 nodes.

MM methods 3 and 4 are based on the assumption that there is a higher probability of locating mobile phone users accurately when identifying nodes of high importance. To determine the importance of a node, the degree centrality measure is used. According to Jiang and Claramunt (2004) and Crucitti et al. (2008) the degree centrality is a topological centrality measure expressing the relationship of a node to its (immediate) neighbors. It can also be considered a local centrality measure, since it only considers the neighborhood, and not the graph as total, as a global centrality measure would (Jiang and Claramunt, 2004; Jiang and Harrie, 2004).

MM method 3 counts the number of adjacent edges of a node. Since the algorithm only considers the immediate neighbor edges, the centrality is called “one-level” degree centrality. MM method 4 additionally considers the number of adjacent edges of the considered node’s neighbor nodes. Since it considers two levels of neighbors, this measure is called “two-level” degree centrality. In both methods, the node with the highest one-level and two-level degree, respectively, is determined as the mobile phone user’s location amongst the candidate nodes. If there are multiple nodes achieving the highest identical degree centrality, the algorithm qualifies the first node of the candidate set as the “most central” one. This introduces a certain degree of randomness. Figure 15 shows the nodes with the highest one-level and two-level degree centralities, respectively.

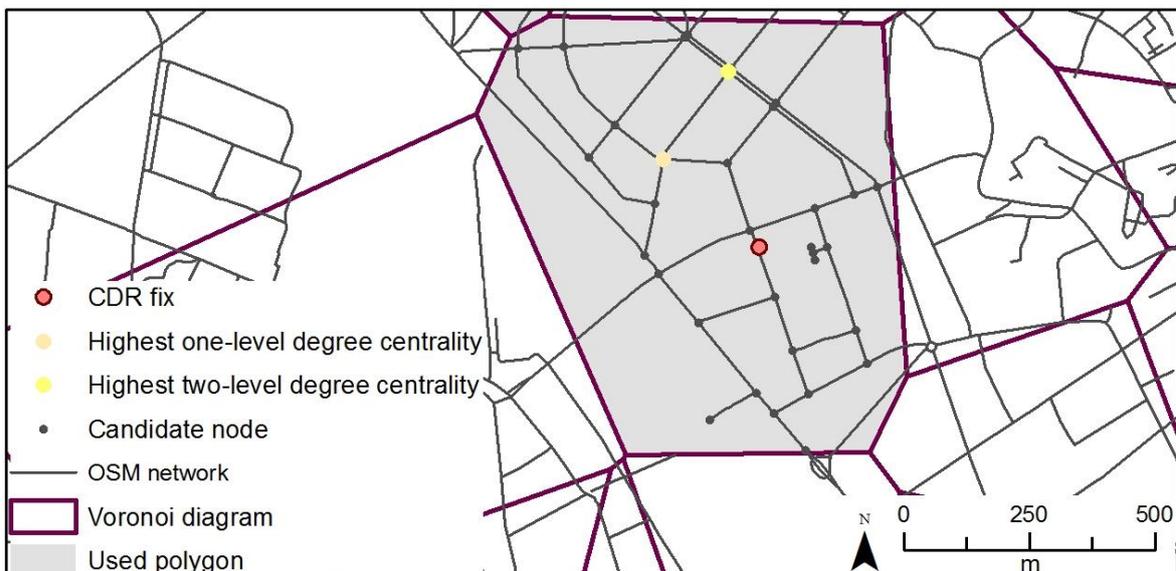


Figure 15: Highest one-level and two-level degree centrality (subset of CDR data of user 4, 03.06.2013)

5.2.5 MM methods 5-7: Approaches relying on edge-based criteria

MM methods 5-7 are based on the identification of candidate edges prior to defining, with the highest probability and based on various assumptions, the edge where the mobile phone user might have been located.

Two-level heuristic to identify the candidate edges

A two-level heuristic to identify the candidate edges, where the user might have been located during a phone activity, is used. In the first step, all edges are chosen as potential edges that lie entirely inside a Voronoi cell. If no edge is entirely contained in a Voronoi cell, then edges that only intersect (are partially contained in) the Voronoi cell are identified as candidate edges. On average, there are 120 entirely contained and 126 intersecting edges per Voronoi cell. There are seven Voronoi cells that do not entirely contain an edge. If, from the beginning, all the edges that intersect the cell would be considered as candidate edges, the case would occur in which an edge that overlaps two consecutively used Voronoi cells would qualify twice as a probable location. In this case, trajectory reconstruction could not be performed between the two respective locations. The decision regarding the identification of candidate nodes based on *intersect* or *contain* operations has consequences for expectations of movement between two neighboring Voronoi cells to happen or not. In this study, it is expected that a movement was expressed by the user when the mobile phone is connected to different neighboring antennas in consecutive phone activities.

MM method 5: Find the edge with the highest road category

MM method 5 is based on the road type, which is stored as an attribute for each edge. The algorithm for MM method 5 iterates through the whole set of candidate edges and retrieves the associated road type of each edge. The road type, which is a nominal value, is subsequently ranked according to its “semantic” importance on the ordinal scale represented in Table 8. The underlying assumption is the following: The higher the score, the more important is the road, and consequently the higher is the probability that the mobile phone user was located on the respective edge.

Table 8: Assigning a ranking to road types

Road type	Ordinal scale
Motorway	11
Trunk	10
Trunk edge	9
Primary	8
Primary edge	7
Secondary	6
Secondary edge	5
Tertiary	4
Tertiary edge	3
Residential	2
Remaining categories	1

MM method 6: Find the edge with the highest speed limitation

The algorithm of MM method 6 works in a similar fashion to MM method 5, except that the considered attribute is speed. The advantage of the speed attribute is that it is a ratio variable. Comparisons between ratio values are feasible without the necessity of a prior ranking, as was the case for the road types. The underlying assumption is again, that edges with higher speed limitations are considered as more important and consequently more frequented roads. Therefore, the probability of an accurate location of a mobile phone user is higher at these edges.

MM method 7: Find the longest edge

In contrast to MM methods 5 and 6, which are based on an edge attribute, MM method 7 is based on a geometric attribute, namely the length of an edge. For MM method 7, all points inside the intersection of the network and the Voronoi cell are assigned an equal probability. This uniform distribution gives no preference to one point over another. With these assumptions, a user is more likely to be found on a longer edge, since it contains more probable positions than a shorter edge. Therefore, in MM method 7 edges are rated according to their length.

Figure 16 shows the identified edges based on the respective edge-based criterion that has been used. As in MM methods 3 and 4, in MM methods 5-7 edges that are in a higher position in the candidate edge set are favored. This is due to the fact that as soon as the first edge reaches the highest value of the respective criterion among the candidate set, further edges which yield equally high values are no longer considered. This again introduces some randomness into the edge selection process. The shortest-path algorithm, described in Section 5.3, only takes nodes as input. For this reason, one of the end nodes of the respective identified edge is considered as the estimated location of the mobile phone user during the phone activity.

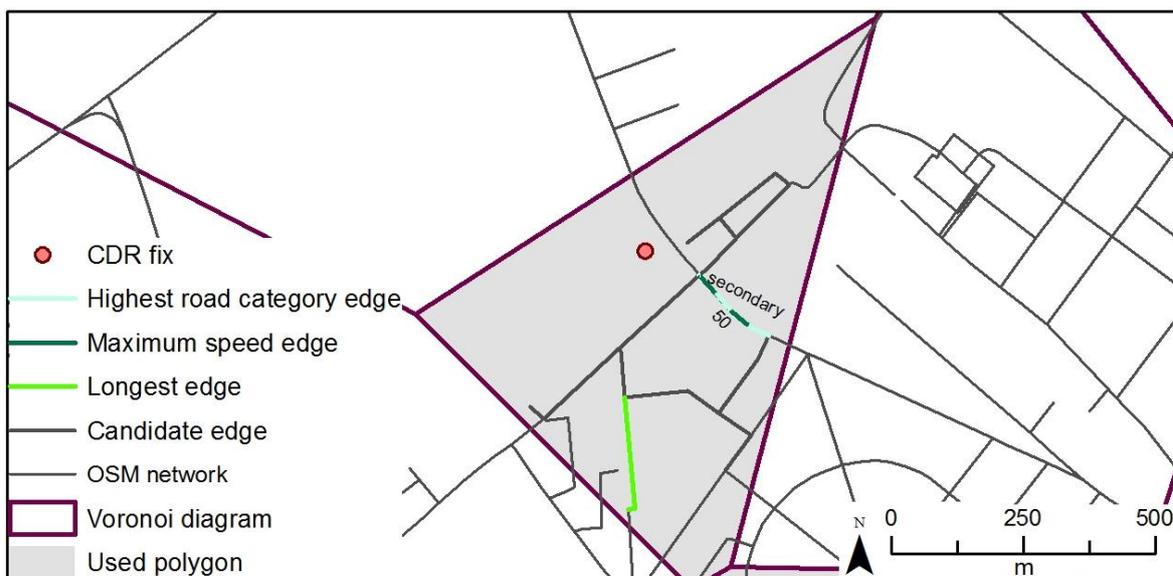


Figure 16: Edges identified on edge-based criteria: the two criteria highest road category and maximum speed identify the same edge, the respective road category (secondary) and the maximum speed (50km/h) are indicated (subset of CDR data of user 4, 03.06.2013)

5.2.6 Pre-validation of CDR map matching

Before the identified graph nodes are used as inputs for the shortest-path computations, a quality estimation for the different map-matching methods is made. As described in Waadt et al. (2009) and Zang et al. (2010), the validation of an estimated position (in our case the node on the network) is done by comparing it to the location of the GPS point at the respective time. Since, in the case of the datasets used here, there is not always a GPS point with corresponding temporal coordinates available, the Euclidean distance between the temporally closest GPS fix and the CDR fix is used as a quality measure for the different map-matching methods. The descriptive statistics in Table 9 and the box plots in Figure 17 show the distance distribution in m for the different MM methods. From an examination of the box plots, it is discernible that the distribution of the distance data appears comparable for all the methods. The minimum distance varies between 5.3 and 26.6 m and the maximum distance between 12'024.1 and 14'858.4 m. The median is between 380.4 and 501.6 m, the mean between 983.5 and 1'802.6 m. The fact that the mean is significantly higher than the median is an indication for a distribution that is skewed to the right. As can be seen in the box plots, there are some important outliers showing high estimation errors (distances). More than 75% of the distances are lower than, depending on the method, 666.6-1'234.0 m.

The validation results of Waadt et al. (2009), who used the center of gravity as the estimated location within the Voronoi cell, show estimation errors below 356 m for 50% of all cases and below 881 m in 90% of all cases. The median value for the equivalent MM method 2 in this study is approximately of the same order of magnitude. The 90th percentile with a value of 4'257.8 m is significantly higher though. The generally high values of the 90th percentile of all the methods are partially caused by the fact that the time gap between the CDR fix and the temporally closest GPS fix is quite large. Statistical analyses showed that for 50% of the cases the closest GPS fix available was below 53 s and for 75% of the cases below 12 min. The maximum time difference was more than 5 h though.

A comparison of the statistics of the different map-matching methods shows that MM method 6, which is based on the speed criterion, – especially when looking at the percentiles – yields considerably lower distance values. 75% of all cases have estimation errors below 971.4 m and 90% of all the cases below 1947.7 m. Therefore, MM method 6 is qualified as the most suitable map-matching method. The comparison of the different methods is further elaborated in Section 6.4.1.

Table 9: Descriptive statistics for the distances in m between CDR and GPS fixes for the MM methods 1-7

	MM 1	MM 2	MM 3	MM 4	MM 5	MM 6	MM 7
Min.	20.7	29.4	17.4	17.4	5.3	5.3	26.6
1st Qu.	181.2	201.7	178.8	246.5	267.8	190.8	194.1
Median	457.9	479.1	462.4	380.4	469.1	412.7	501.6
Mean	1252.3	1389.5	1802.6	1765.8	1423.8	983.5	1196.8
3rd Qu.	934.8	1234.0	1053.6	1022.6	971.4	666.6	1149.9
.9 Perc.	3882.8	4257.8	7533.3	8201.7	4507.6	1947.7	2798.2
Max.	13954.6	12141.3	14858.4	12024.1	12442.5	12024.1	10396.1

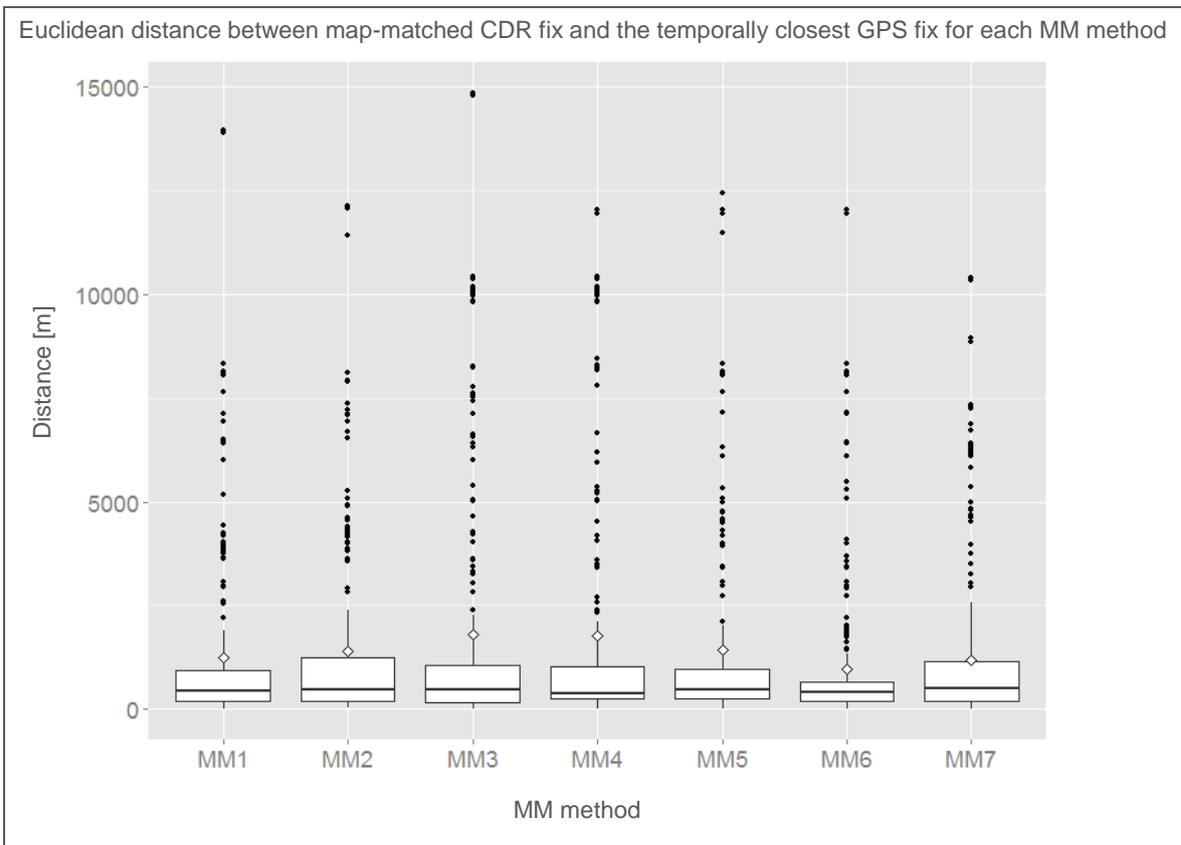


Figure 17: Box plots showing the Euclidean distances between the map-matched CDR fixes (according to MM methods 1-7) and the temporally closest GPS fixes, the diamond represents the mean

5.3 Shortest path between the selected nodes

As soon as the CDR data is assigned to a node on the network by one of the above-described map-matching methods, a shortest-path algorithm is implemented to find the shortest connections between consecutive locations on the network (Wentz et al., 2003). Although there are other approaches to compute the way between two locations, such as simplest path proposed by Duckham and Kulik (2003), shortest-path is still the most frequently used one in navigational research (Fu et al., 2006; Järv et al., 2012). Since computational efficiency is not primarily of importance, Dijkstra (1959) shortest-path algorithm, which is implemented in GeoTools (2014), is used to identify the edges that have been travelled between two different locations. The algorithm computes the shortest paths from

a single node to all the other nodes on a graph. Since the shortest path for a whole sequence of nodes needs to be computed, the GeoTools *DijkstraShortestPathFinder*¹¹ (DSPF) needs to be adapted in a way that allows sequential shortest-path computations. Also, a filter condition must be implemented which guarantees that consecutive nodes with identical location are only considered once in the shortest-path computations.

Before the shortest path can be computed, a weighting strategy for the graph must be defined. In contrast to Jiang and Claramunt (2004), who use an unweighted graph where each edge has a unit distance, in this study, the edges are weighted according to their lengths. The *EdgeWeighter*¹² interface of GeoTools (2014) is therefore used. The sequential shortest-paths algorithm works in the following way:

Algorithm: Sequential shortest-paths computation

Input:

```
NodeSet nodeSet //map matched CDR locations
```

Initialize:

```
Graph graph
```

```
FeatureCollection shortestPathCollection //feature collection where shortest path is stored
```

```
//define an edge weighting strategy (edges are weighted according to their lengths)
```

```
EdgeWeighter weighter = new EdgeWeighter()
```

```
    getWeight(Edge edge) //edge as input of getWeight method
```

```
        return edge.getLength() //output is the length of the edge
```

```
//calculate shortest paths between the sequence of nodes
```

```
for (int i = 1; i<nodeSet.size; i++)
```

```
    Node sourceNode = nodeSet.get(i-1)
```

```
    Node destinationNode = nodeSet.get(i)
```

```
    if(!sourceNode.equals(destinationNode)) //test locations are not identical
```

```
        DijkstraShortestPathFinder dspf = new DijkstraShortestPath
```

```
        Finder(graph,sourceNode,weighter) //initialization of the Dijkstra
```

```
        ShortestPathFinder using the arguments graph, sourceNode and weighter
```

```
        dspf.calculate() //shortest-path computation from source to all graph nodes
```

```
        ArrayList<Edges> pathEdges = dspf.getPath(destinationNode) //path
```

```
        from source node to destination is retrieved
```

```
        if (pathEdges.size=0)
```

```
            System.out.println("path to that destination could not be computed →probably an error with the network dataset")
```

```
        for (int i; i<pathEdges.size(); i++ //iterate through path
```

```
            shortestPathCollection.add(pathEdges.get(i)) //adding shortest path of the first source - destination couple
```

As observable in the algorithm, the DSPF needs to be reinitialized for every source – destination couple, since the shortest path to all nodes for only one source node is computed each time. The output message, that a path to a destination could not be computed due to a problem with the network dataset, showed up only for one set of input nodes – the daily

¹¹ <http://udig.refractions.net/files/docs/api-geotools/org/geotools/graph/path/DijkstraShortestPathFinder.html> (accessed 13.5.2014)

¹² <http://udig.refractions.net/files/docs/api-geotools/org/geotools/graph/traverse/standard/Dijkstralterator.EdgeWeighter.html> (accessed 13.5.2014)

segment 6_130605 map matched by MM method 4 (center of gravity rationale). The respective daily segment was therefore excluded from the data sample, which is thereby reduced from 81 to 80 cases. Figure 18 shows the visualization of the reconstructed paths based on MM method 6 (maximum speed rationale) and MM method 7 (longest edge rationale). When comparing the reconstructed paths to the GPS points in this map, it seems that the MM method 6 based reconstructed path better imitates the ground truth data. Quantification of the similarity between a reconstructed path and the ground truth will be subject in the following Chapter 6.

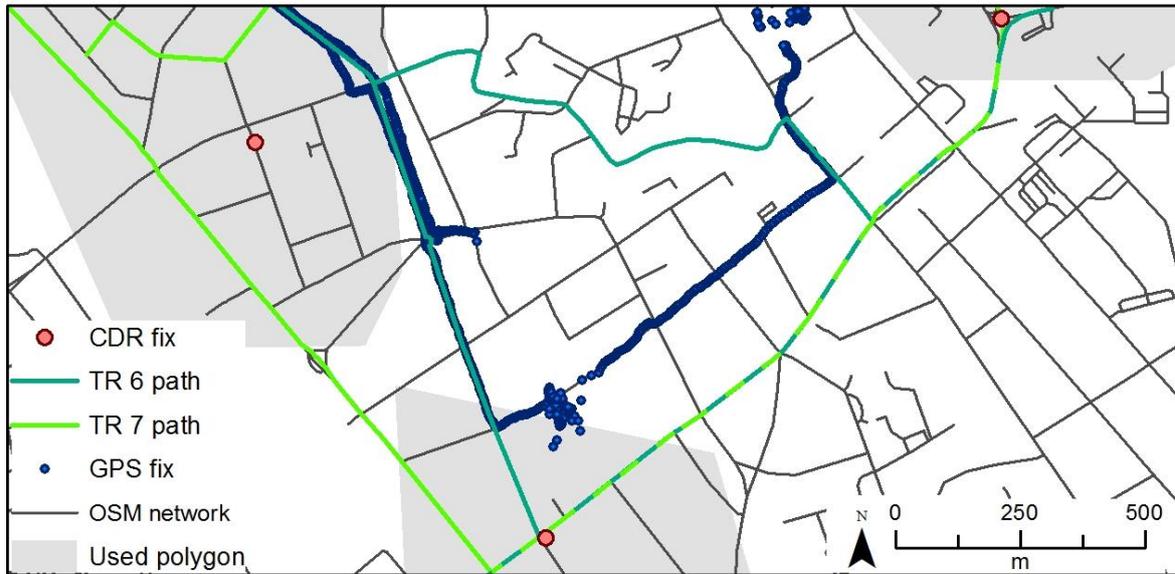


Figure 18: Reconstructed trajectories on the basis of MM method 6 (TR 6, maximum speed rationale) and MM method 7 (TR 7, longest edge rationale) (subset of CDR data of user 4, 03.06.2013)

5.4 Summary of trajectory reconstruction methods

Table 10 shows an overview of the seven different trajectory reconstruction (TR) methods that have been developed in this study. In a first step, the CDR fixes have been map matched to a node of the network by application of one of the seven different rationales presented in Sections 5.2.1 - 5.2.5. In a second step, the sequential shortest-paths algorithm – described in Section 5.3 – is applied, to find the shortest connections between the identified nodes.

Table 10: Overview of trajectory reconstruction (TR) methods and the respective underlying map-matching (MM) methods

TR method	MM Rationale
TR 1 (based on MM 1)	Closest node to antenna
TR 2 (based on MM 2)	Closest node to center of gravity of Voronoi cell
TR 3 (based on MM 3)	Highest one-level degree centrality node
TR 4 (based on MM 4)	Highest two-level degree centrality node
TR 5 (based on MM 5)	Highest road category
TR 6 (based on MM 6)	Maximum speed edge
TR 7 (based on MM 7)	Longest edge

6 Validation

6.1 Overview

The validation of the developed trajectory reconstruction methods is documented in this chapter. Figure 19 is a visualization of the different steps that lead to a quantification of the similarity between the reconstructed paths and the ground truth.

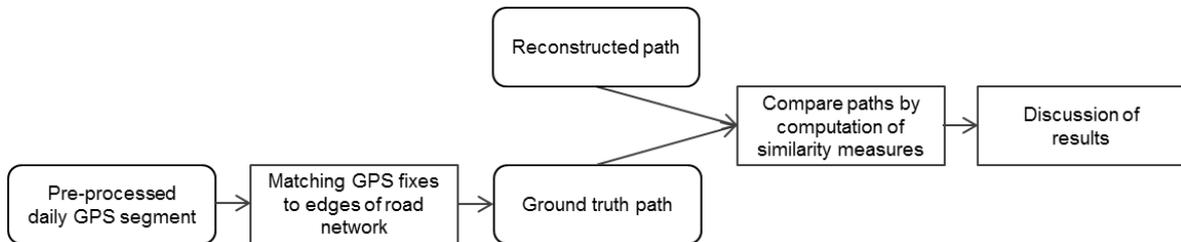


Figure 19: Workflow of validation of trajectory reconstruction methods by comparing reconstructed paths to the ground truth

To validate the trajectory reconstruction methods, the reconstructed paths are compared to the GPS points that constitute the ground truth. Therefore, the GPS points need to be converted to units comparable to the reconstructed paths. This is accomplished by identifying the edges representing the travelled path, through matching the GPS points to the road network (Section 6.2). Similarity measures are computed (Section 6.3) and statistically analyzed (Section 6.4) to compare the resulting ground truth paths to the reconstructed paths.

6.2 Making paths from GPS fixes

The most likely travelled on graph edges are identified and stored as a set of edges in order to obtain a ground truth that is comparable to the reconstructed paths. Therefore, the following three steps are undertaken:

1. Identify all the edges that are closest to at least one GPS fix based on the Euclidean distance between the GPS fix and its perpendicular projection to the edge.
2. Disqualify unlikely edges based on an edge-score criterion (number of GPS points projected to a specific edge in relation to its length).
3. Fill the gaps by application of a shortest-path heuristic to construct a continuous ground truth path.

The three above-mentioned steps are described in detail in Sections 6.2.1 - 6.2.3 and visualized in Figure 20. In a last step, the ground truth paths are compared to the original GPS data in order to assess the quality of the ground truth path. If a path deviates considerably from the corresponding original data, it is disqualified. The respective filtering criterion is described in Section 6.2.4.

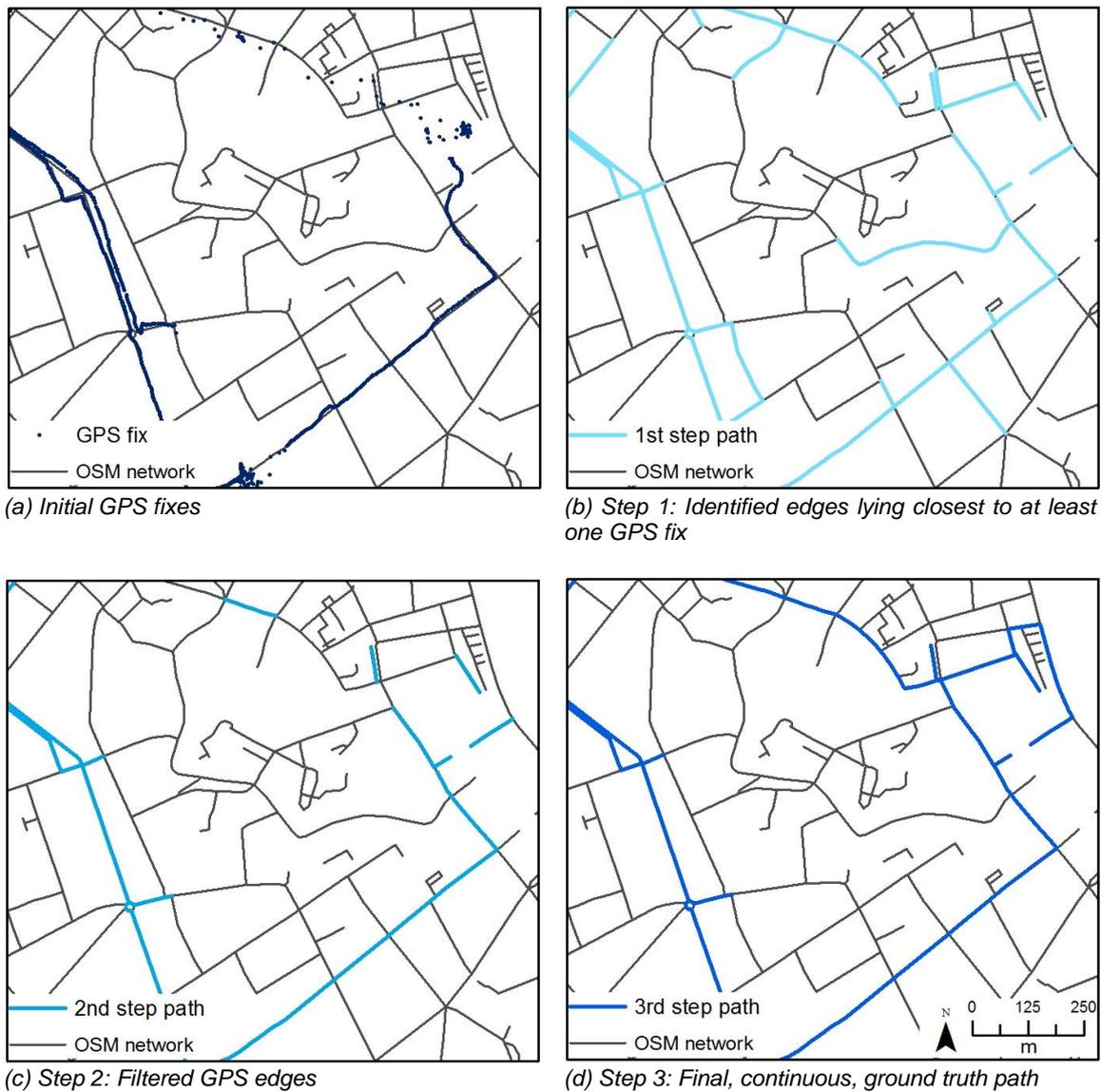


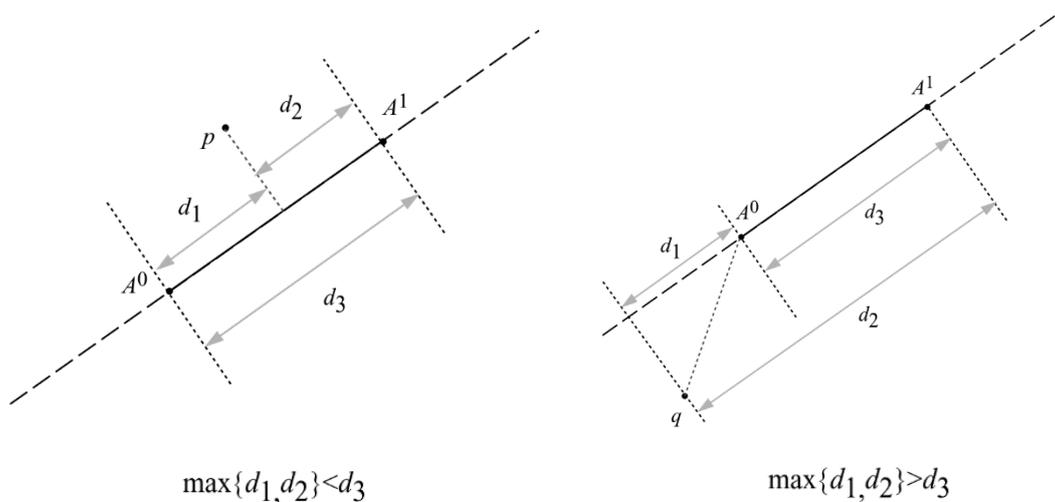
Figure 20: Visualization of initial GPS fixes and the three steps for deriving the continuous GPS path (subset of GPS data of user 4, 03.06.2013)

6.2.1 1st step: Identifying edges closest to GPS fixes

In the first step, the raw positioning points (Figure 20a) need to be matched to one of the edges of the road network (Smoreda et al., 2013). This is done by defining the edge located closest to the GPS point, based on the Euclidean distance between the GPS point and the edges. The thereby identified edges are consolidated into a set of edges maintaining the order of the movement. An edge, however, is only added to the set, if it is not already included. Therefore, multiple travelled paths are not distinguishable.

The algorithm follows the same principles as algorithm 1 used to map match points to an arc described in White et al. (2000, p. 96). The computation of the distance between a point and a line segment depends on whether the point is perpendicularly projectable to the line segment (as in Figure 21(a)) or not (as in Figure 21(b)). In the first case, the distance between point and line segment equals the Euclidean distance between point p and its perpendicular projection p' to the line segment. In the latter case, this would be the Euclidean distance between point q and its projection to the start or end point q' of the line segment –

favoring the shorter one of the two. To project the point to the line segments the *project*¹³ function of *LocationIndexedLine* provided by GeoTools (2014) is used.



(a) Point p is perpendicularly projectable to line segment if $\max\{d_1, d_2\} < d_3 = \text{true}$

(b) Point q is not perpendicularly projectable to line segment if $\max\{d_1, d_2\} > d_3 = \text{true}$

Figure 21: Distance calculation between a point and a line segment defined by A^0 and A^1 (White et al., 2000)

As observed in Figure 20(b), there is one major constraint to this straightforward approach. As also stated by White et al. (2000), edges, which obviously have not been traveled when looking at the sequence of GPS points, are identified. This is particularly problematic at junctions where an adjacent (certainly not travelled on) edge happens to be located closest to a GPS fix due to imprecise GPS measurements. The removal of these unintentionally identified sideways by the first step approach is documented in the following Section 6.2.2.

6.2.2 2nd step: Removing unintentionally identified edges

Edges which were unintentionally identified are discarded based on a minimum required edge-score value, which is computed for each edge by counting the number of GPS points matched to it divided by its length. A comparable approach is used by Smoreda et al. (2013), who classify candidate edges as “ambiguous” based on the number of CDR fixes that are matched to an edge. The threshold value of the edge scores is empirically defined by investigation of several GPS daily segments: Firstly, the number of edges that should be eliminated e is counted by visual comparison of the original GPS points and the identified edges of the 1st step path. Subsequently, the edge scores of all the edges composing the 1st step path are computed and listed in descending order. Thirdly, the edge score at the e^{th} + 10% last position, which represents the appropriate threshold value for the investigated data unit, is retained. The increase of e by 10% assures that with a higher probability all the unintentionally identified edges are removed. Repeating this for 10 randomly selected data segments¹⁴, a mean edge-score value of approximately 0.035 was obtained for the ground

¹³ <http://www.vividsolutions.com/jts/javadoc/com/vividsolutions/jts/linearref/LocationIndexedLine.html> (accessed 13.5.2014)

¹⁴ The threshold value of the edge-score computation is based on the following GPS daily segments: 1_130618, 1_130625, 3_130621, 4_130602, 4_130622, 5_130602, 5_130625, 6_130618, 6_130620, 7_130627

truth data. The standard deviation is 0.02, indicating a rather small variance, which again confirms that the threshold value seems to be representative for the investigated dataset.

In the 2nd step, edges which were identified in the first step are filtered retaining only edges that have an edge-score value lower than 0.035. Since the GPS signal might be weak in certain areas, edge scores might lead to a disqualification of edges which obviously should be part of the ground truth movement. In Figure 20(c), this phenomenon is observable, especially in the north-eastern part of the 2nd step path. This problem is tackled in the following Section 6.2.3.

6.2.3 3rd step: Making GPS path continuous

Since it is reasonable to assume that a travelled path is continuous, the occurring gaps in the 2nd step path – due to weak or temporarily missing GPS signal – are refilled in the third step. The gap refilling is based on the shortest-path heuristic, as described in Section 5.3. The following steps lead to a continuous path: A gap between two consecutive edges is identified, if the smallest distance between the two of them is greater than 0. Subsequently, the shortest paths for the four possible combinations of start and end node of the two edges are computed. The path with the smallest number of edges among the four shortest paths is used as gap filler and is thus added to the filtered edge set that resulted from step 2. The final ground truth path is visualized in Figure 20(d). As a result of step 3, the sequence of the edges is no longer maintained. Therefore, the resulting ground truth path is purely spatial. The only temporal information available is that the movement took place within the same time period (minus max. 30 min) as the available CDR data, as a result of the initial data clipping (see Chapter 4.3.3). Since the GPS map matching includes shortest-path computation, problems with the road network data might occur (cf. Chapter 5.3). This was the case for two GPS files¹⁵, which were consequently excluded from the data sample, which is thereby reduced from 80 to 78 analyzable segments.

6.2.4 Disqualification of unsuitable ground truth paths

A measure that works as proxy for closeness of a path to the original GPS fixes is applied, disqualifying paths that were poorly matched to the road network (due to noisy GPS signals or methodological limitations) from serving as ground truth. The measure is computed by taking the average of all distances from the path edges to their respective closest original GPS points. The diagram in Figure 22 shows this measure for each ground truth path in ascending order. A clear change of slope is indicated with an orange line at a value of 54 m. This value will serve as maximum tolerable average distance of the path to the original data. Applying this threshold value, 5 paths out of the 78 remaining GPS paths are excluded.

¹⁵ The two daily segments 1_130621 and 4_130621 are therefore excluded for further analysis.

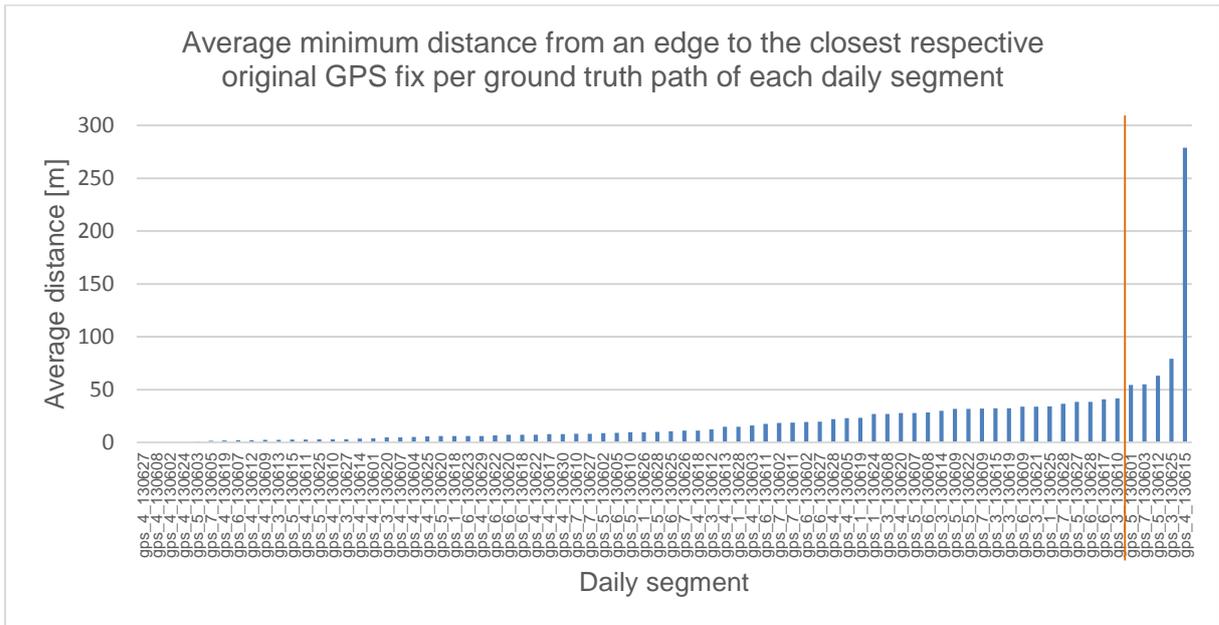


Figure 22: Diagram of average distance from an edge to the closest GPS fix per ground truth path

6.3 Assessing similarity between reconstructed and ground truth paths

6.3.1 Units to be compared

Eventually, for each of the 73 remaining daily segments one ground truth path as well as seven paths reconstructed in different ways are available. By comparison of the paths reconstructed by one of the seven different TR methods to the ground truth paths, an operational validation in the sense of Rykiel (1996) of the respective TR method is done. Figure 23 shows an exemplary constellation of a ground truth path and a path reconstructed with TR method 4 of the daily segment 4_130603.

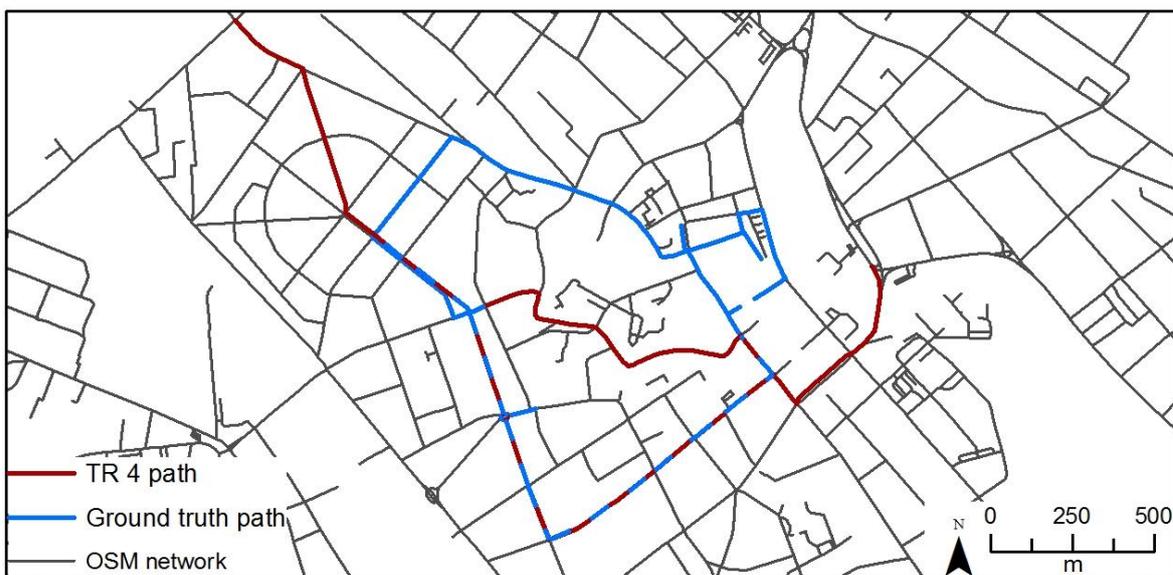


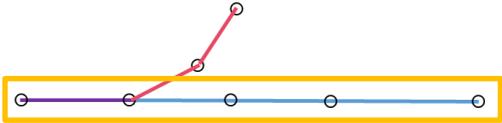
Figure 23: Path reconstructed by TR method 4 and corresponding ground truth path, allowing quantification of the similarity of the two paths (based on CDR and GPS data of user 4, 03.06.2013)

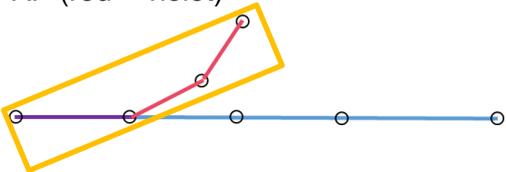
6.3.2 Computation of similarity measures

Similarity measures need to be defined in order to assess the similarity between two paths. Since the developed ground truth paths do not contain any temporal – or sequential – information, the focus is put on spatial similarity measures. Since no multiple travelled paths are considered in the ground truth, duplicate (in terms of spatial equality) edges are removed from the reconstructed paths, before similarity computations are performed.

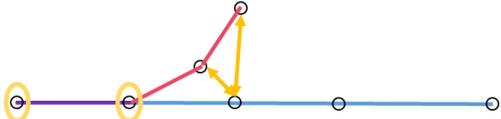
Since there is no universally suitable similarity measure, a range of similarity measures are implemented and tested (Brakatsoulas et al., 2005; Wentz et al., 2003). The similarity is computed pairwise for each reconstructed and ground truth path couple. Depending on the specific question at hand, the fulfilling of different similarity criteria might be of interest and therefore, the measure that most appropriately reflects these criteria can be used. The implemented similarity measures (SM) and their advantages and disadvantages are listed in Table 11. It must be noted, however, that certain characteristics of the SMs which are listed on the advantage side, might equally be listed amongst the disadvantages, and vice versa. The similarity measures which do not automatically yield values between 0 and 1, are normalized to a range between 0 and 1, in order to be mutually comparable. Higher values indicate higher similarities, whereas 0 means no similarity at all and 1 signifies perfect similarity in terms of the respective SM.

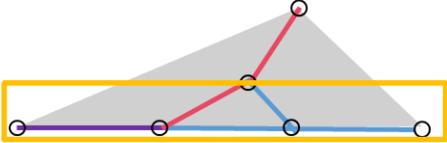
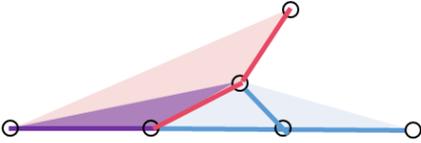
Table 11: Overview of the different proposed and implemented similarity measures (SMs)

SM	Approach	Advantage	Disadvantage
<i>Comparison of total length / number of edges of the paths</i>			
SM1	Ratio of the number (<i>no.</i>) of edges (<i>E</i>) of the longer path (in terms of <i>no.</i> of <i>E</i>) to the <i>no.</i> of <i>E</i> of the shorter path. In the sketch: <i>no.</i> of <i>E</i> of reconstructed path (<i>RP</i>) (red) divided by <i>no.</i> of <i>E</i> of ground truth path (<i>GTP</i>) (blue)	– Easy to implement and fast (a)	– If value close to 1, no guarantee that paths are spatially similar (b)
			
SM2	Ratio of the length (<i>L</i>) of the longer path to the <i>L</i> of the shorter path. In the sketch: <i>L</i> of <i>RP</i> (red) divided by <i>L</i> of <i>GTP</i> (blue)	– (a) – Edges weighted according to their length (c)	– (b)
			
<i>Similarity measures based on alignment in terms of edges</i>			
SM3	Ratio of the <i>no.</i> of shared <i>E</i> between <i>RP</i> and <i>GTP</i> (violet) to the total <i>no.</i> of <i>E</i> of the <i>GTP</i> (blue + violet)	– Exact and strict measure (d) – Measure of how well ground truth path is approximated by reconstructed path (e)	– Problematic if paths are of different lengths (f) – Boolean criteria: Two paths lying close together, but following parallel edges, are not considered similar at all (g)
			

SM	Approach	Advantage	Disadvantage
SM4	Ratio of L of shared E between RP and GTP (violet) to the total L of the GTP (blue + violet)	– (c) – (d) – (e)	– (f) – (g)
			
SM5	Ratio of no. of shared E between RP and GTP (violet) to the total no. of E of the RP (red + violet)	– (d) – Measure of how much of the reconstructed path was actually travelled (h)	– (f) – (g)
			
SM6	Ratio of L of shared E between RP and GTP (violet) to the total L of the RP (red)	– (c) – (d) – (h)	– (f) – (g)
			

Normalized Hausdorff distance based similarity measures

SM7	Bidirectional Hausdorff distance $D_H(GTP, RP)$ as defined in Alt and Guibas (1996): The longest shortest distance from any node of the GTP (blue) to its closest node on the RP (red), or vice versa from RP to GTP (distance in yellow)	– Does not account for exact alignment, but looks more at an overall closeness (i)	– (f) – Susceptible to outliers (j) – Difficult normalization procedure (k)
			
	Normalization: If $D_H(GTP, RP) > L$ of $GTP \rightarrow SM7 = 0$; otherwise the $D_H(GTP, RP)$ is divided by the L of the GTP , and this fraction is subsequently subtracted from 1.		
SM8	Average (avg.) directed Hausdorff distance from RP ($avg. \vec{D}_H(RP, GTP)$): Avg. distance to travel from a node of the RP (red) to the closest node on the GTP (blue) (averaged L of yellow arrows)	– (h) – (i) – Less susceptible to outliers (l)	– (f) – (k)
			

SM	Approach	Advantage	Disadvantage
	Normalization: If $avg. \vec{D}_H(RP, GTP) > 1000 \text{ m}^{16} \rightarrow SM8 = 0$; otherwise the $avg. \vec{D}_H(RP, GTP)$ is divided by 1000 m, and this fraction is subsequently subtracted from 1.		
SM9 ¹⁷	Avg. directed Hausdorff distance from GTP ($avg. \vec{D}_H(GTP, RP)$): Avg. distance to travel from a node of the GTP to the closest point on the RP (averaged L of yellow arrows)	<ul style="list-style-type: none"> – (e) – (i) – (l) 	<ul style="list-style-type: none"> – (f) – (k)
	 <p>Normalization: Normalization procedure is identical to that of SM8.</p>		
<i>Similarity measures based on convex hull of the paths</i>			
SM11	Ratio of the area of the convex hull of the combination of GTP and RP (grey) to the total L of the GTP (blue + violet)	<ul style="list-style-type: none"> – (i) – By taking into account length of ground truth paths, long or roundish paths are treated equally (m) 	<ul style="list-style-type: none"> – (f) – Convex hull is a quite rough estimation of the area that has been covered (n)
	 <p>Normalization: Normalization procedure is identical to that of SM8.</p>		
SM12	Ratio of the area of intersection of the convex hull of the GTP and the convex hull of the RP (light violet) to the area of the union of the two separate convex hulls (light red, light blue + light violet)	<ul style="list-style-type: none"> – (i) – Measure reflects whether area of movement described by the two paths coincides (o) 	<ul style="list-style-type: none"> – (n) – Sensitive to extreme cases of two paths that are identical and straight lines except for the last edge, where the two paths go into opposite directions, resulting in no overlap of the convex hulls (p)
			

¹⁶ The threshold of 1000 m was determined by investigation of the distribution of the values of average Hausdorff distances from the reconstructed paths of all comparison cases in ascending order. The threshold is set to the value where the most significant change of slope in the distribution is observed.

¹⁷ SM10 was rejected. For this reason the numeration is not continuous.

6.3.3 Discussion of similarity measures

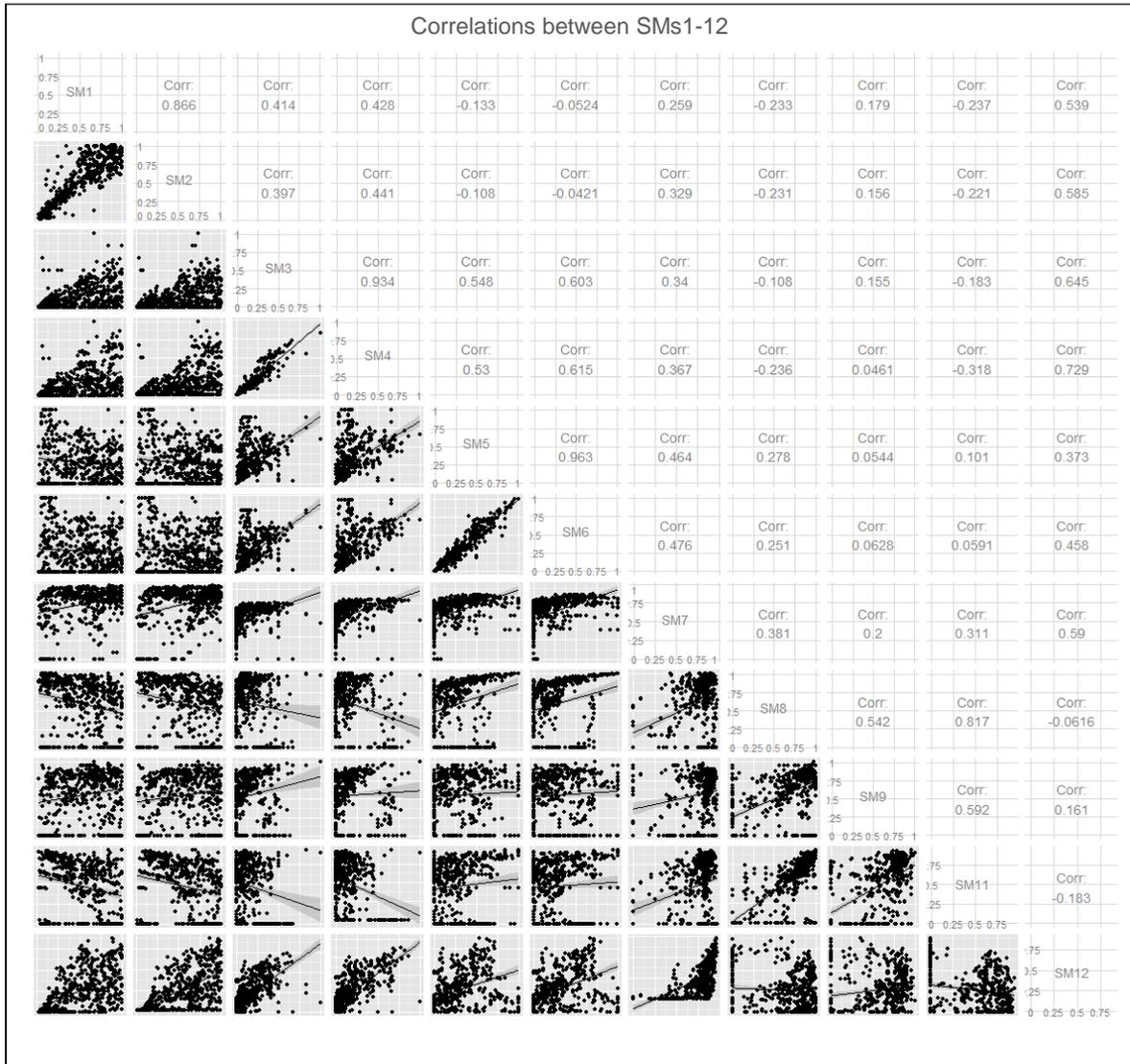


Figure 24: Scatter plot matrix for the different similarity measures including correlation values for all the 511 comparison cases

Figure 24 shows a scatter plot matrix including the correlation values for the 11 similarity measures that have been computed for all the 511 comparison cases¹⁸. It is notable that the SMs 1 and 2, 3 and 4, as well as 5 and 6 show high correlation values of 0.866, 0.934 and 0.963, respectively. This is unsurprising, since the only difference between each pair of SMs is that either the number of the edges or the actual length of the edges is considered. In contrast to the SMs 3-6 that capture similarity based on the exact alignment of the edges of two paths and therefore are quite strict measures, SMs 1 and 2 do not account for the spatial arrangement at all. SMs 1 and 2 are basically useful to give a first impression whether two paths are within the same range in terms of scale of the movement.

The Hausdorff distance based similarity measures (SMs 7-9) introduce an interesting aspect into the quantification of path similarity. In contrast to the Boolean SMs 3-6, which check whether the alignment of edges is identical or not, the Hausdorff distance based SMs

¹⁸ The application of seven different trajectory reconstruction methods to the 73 different daily CDR segments results in 511 different reconstructed paths that are compared to the corresponding ground truth path.

as well as the SMs 11 and 12 have the capability to make a more differentiated statement about the overall closeness between two paths that do not necessarily follow the same edges. The main constraint of SMs 7-9, however, is that there is no obvious way to normalize them. The normalization approaches which are tested in this study led to little convincing distributions of similarities. It is obvious from the plots for SMs 7-9 in Figure 24 that many values are equal to 0 and subsequently quite rapidly augment to values close to 1. The Hausdorff distance based SMs would be better suited to compare paths that are mutually comparable, especially in terms of total path length. For example, the non-normalized SMs 7-9 might give a very interesting indication about which TR method works best for one specific daily segment.

SM 11, which is amongst the convex hull based similarity measures, has the same issue with normalization as have the Hausdorff based distances. What is remarkable though, are the rather high correlation values of 0.817 and 0.592 between SMs 11 and 8, and SMs 11 and 9, respectively. This indicates that the average Hausdorff distance based measures and the SM 11 seem to comparably capture path similarity, although they are computed in different ways. By dividing the whole area covered by the two paths (approximated by the convex hull of both paths) and subsequently dividing it by the length of the ground truth path, the SM 11 is a proxy for the average distance between the two paths, which is also the case for SMs 8 and 9. This again explains the above-mentioned correlations between the two SMs. SM 11 can be seen as a straightforward approach of the LIP similarity measure proposed by Pelekis et al. (2011), who use the area between two trajectories as distance measure between them (see Figure 2, Section 2.4.1).

What applies for SMs 11 and 12 is that the convex hull is a rather rough description of the actual area covered by the movement. A concave hull would be a more meaningful representation. However, there is no exact definition of the concave hull and the therefore available algorithms are rather sophisticated. For example, when using the alpha-shape algorithm of Wei (2008), the parameter alpha, which controls the precision of the boundary, needs to be specified. This parameter is very dependent on the point density (path nodes) and the scale of the movement. And since these two parameters vary considerably over the whole set of available paths, the parameter would need to be adjusted individually for the different cases. The reasons which led to the use of the convex hull instead of the concave hull are the clear definition for the convex hull for a set of points (or a polyline or polygon which are decomposable to a set of points) (Avis et al., 1997; Efron, 1965) as well as the available implementation¹⁹ on GeoTools (2014).

6.3.4 Selection of similarity measures

Since SMs 1 and 2 do not take into account the spatial arrangement of the two paths, they are not further considered in the analysis of the results. Amongst the similarity measures based on alignment in terms of edges (SM 3-6), a focus is put on SM 4, since it weighs edges according to their length and uses the ground truth path as reference that should be approximated as closely as possible. As in Newson and Krumm (2009), the length of the

¹⁹ <http://tsusiatsoftware.net/jts/javadoc/com/vividsolutions/jts/geom/Geometry.html#convexHull%28%29> (accessed 14.5.2014)

ground truth is used as reference. But while Newson and Krumm compute the total error (ratio of the sum of erroneously identified and not identified edges to the total length of the true path), in this case the ratio of the number of correctly identified edges to the total length of the ground truth path is computed.

Since no appropriate way was found to normalize the set of Hausdorff distance based similarity measures in order to make the measures comparable to the remaining similarity measures, they are not used in further analyses. The same issue regarding normalization applies for SM 11 which is amongst the two convex hull based similarity measures. SM 12 (the second of the convex hull based similarity measures), which puts the shared area of the intersection of the separate convex hulls of the ground truth path and the reconstructed path in ratio with the union of the two convex hulls, has the advantageous property of automatically yielding values between 0 and 1.

Therefore, SM 12 is the second SM – next to SM 4 – that is used for further analysis. By comparison of the convex hulls of the separate paths, it is more tolerable than SM 4, and basically testifies to what extent the areas of movement are identical. The fact that SM 4 is in ratio to the length of the ground truth path only, and SM 12 in ratio to the area of the union of the convex hulls of both paths, leads to the result that SM 12 yields values in a comparable range, although the criterion is less strict. Additionally, the high correlation value for SMs 4 and 12 of 0.729 (cf. Figure 24) indicates that the two measures seem to perceive the paths in a comparable way in terms of similarity.

6.4 Results

As mentioned above, SM 4 and SM 12 are used to analyze the different trajectory reconstruction methods in Section 6.4.1, as well as to analyze possible factors having an impact on the quality of the resulting reconstructed paths in Section 6.4.2. Thus, all analyses in the following are based on the assumption that SMs 4 and 12 are capable of adequately capturing the expected similarities between paths.

6.4.1 Comparison of different trajectory reconstruction algorithms

Table 12 and Table 13 give an overview of the distribution of the SMs 4 and 12 for the different TR methods, respectively, as well as the box plots for the two SMs for each method in Figure 26. The mean and the median lie within a rather small range for the different TR methods, namely 0.15-0.21 for SM 4 and 0.23-0.29 for SM 12, and 0.07-0.16 for SM 4 and 0.19-0.26 for SM 12, respectively. These – rather low – values indicate that the TR methods generally do not seem to perform very well. The 3rd quartiles are not much more promising. For example, 75% of the paths reconstructed with TR method 1, which is approximately in the mid-range of the proposed methods, reach a SM 4 lower than 0.24 and a SM 12 lower than 0.41.

Figure 26 demonstrates that the paths generally obtain higher similarity values with SM 12 than with SM 4. This was to be expected since SM 12 is comparing the shared area of movement between the two paths, whereas SM 4 is comparing the shared edges between

the two paths, which is a much stricter criterion. The box plots also show that the distributions of the similarity values are comparable for the different methods, but that slight differences in the scores for each method are observable. Particularly in terms of SM 4 but also in terms of SM 12, TR method 6 receives the highest scores. The conclusions so far are congruent with the pre-validation of the map-matching methods in Section 5.2.6, where likewise only small differences between the methods could be assessed, but MM method 6 showed the best performance. The worst similarities are obtained by TR method 3 (which was based on the one-level degree centrality). In general, the results suggest that the TR methods relying on edge characteristics generally perform slightly better. The evidence for this statement is rather fragile though. Since MM method 6 produces the most reliable results, the analysis is mostly focused on MM method 6 in the following part of the analyses.

Also the scatter plot in Figure 25, whose dots represent the values of SM 4 and SM 12 for the 511 different paths that have been color-coded according to their TR algorithm, gives no evidence for an unambiguously outperforming algorithm. With some goodwill one might argue that most of the dots in the upper right part of the plot are in the color of TR method 6. A clear pattern, however, is not discernible.

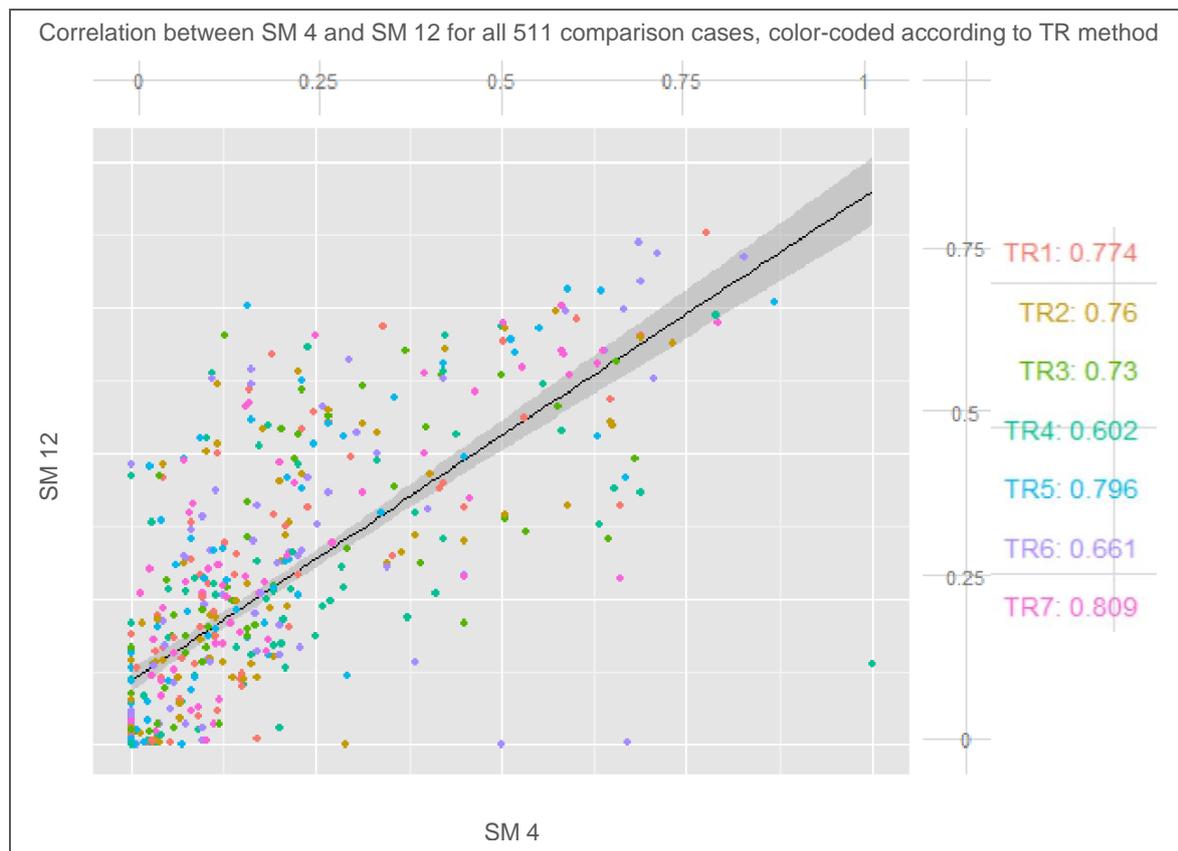


Figure 25: Scatter plot for SMs 4 and 12 for all 511 comparison cases, color-coded according to the TR method

Table 12: Descriptive statistics for SM 4 for the different trajectory reconstruction methods

	TR 1	TR 2	TR 3	TR 4	TR 5	TR 6	TR 7
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1st Qu.	0.03	0.03	0.00	0.02	0.00	0.04	0.03
Median	0.12	0.12	0.07	0.15	0.09	0.16	0.12
Mean	0.18	0.18	0.15	0.20	0.16	0.21	0.19
3rd Qu.	0.24	0.23	0.22	0.27	0.23	0.30	0.22
Max.	0.78	0.73	0.68	1.00	0.87	0.83	0.79

Table 13: Descriptive statistics of SM 12 for the different trajectory reconstruction methods

	TR 1	TR 2	TR 3	TR 4	TR 5	TR 6	TR 7
Min.	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1st Qu.	0.03	0.06	0.02	0.03	0.03	0.05	0.06
Median	0.20	0.21	0.19	0.22	0.21	0.26	0.26
Mean	0.26	0.26	0.23	0.25	0.27	0.29	0.28
3rd Qu.	0.41	0.45	0.39	0.38	0.48	0.44	0.43
Max.	0.88	0.74	0.70	0.74	0.78	0.86	0.75

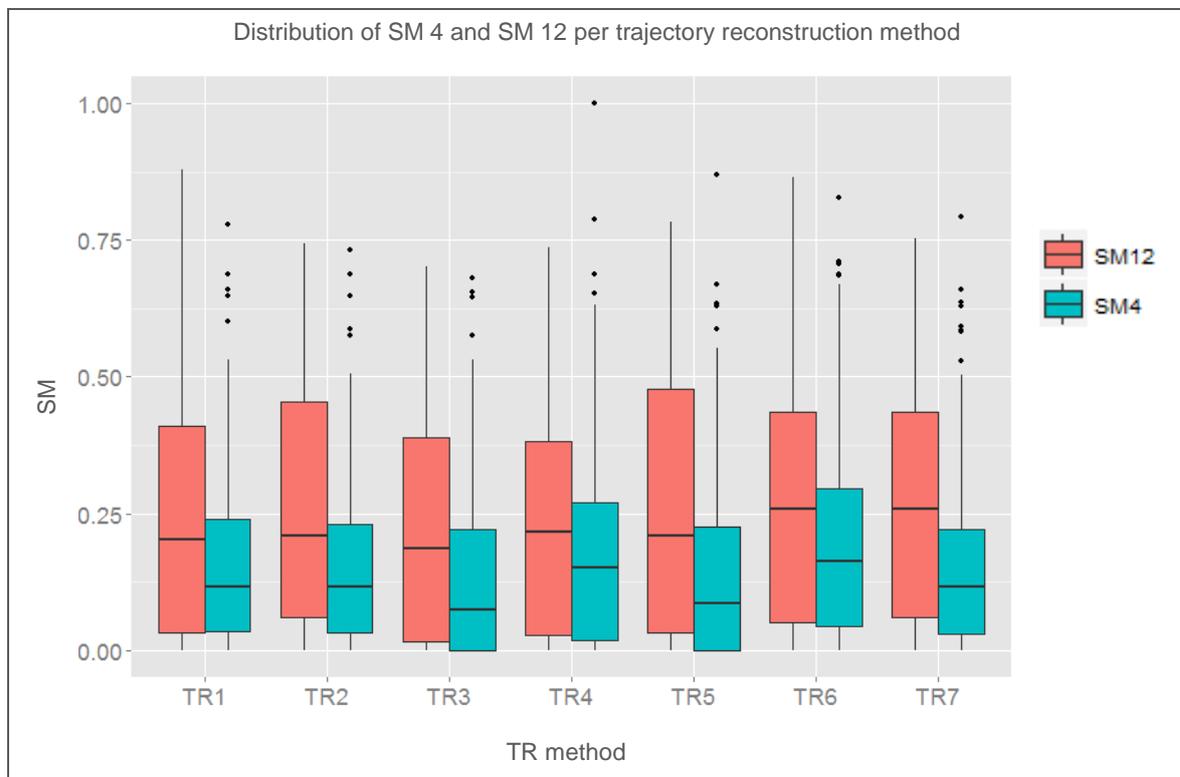
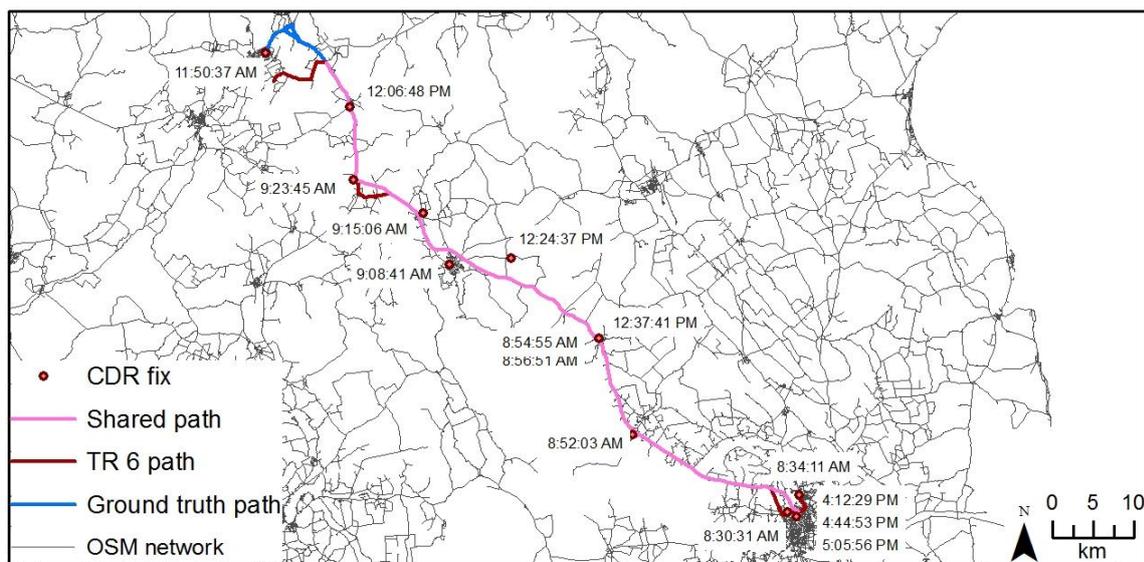


Figure 26: Box plots of the values of SMs 4 and 12 for the 73 reconstructed daily paths with the different TR methods

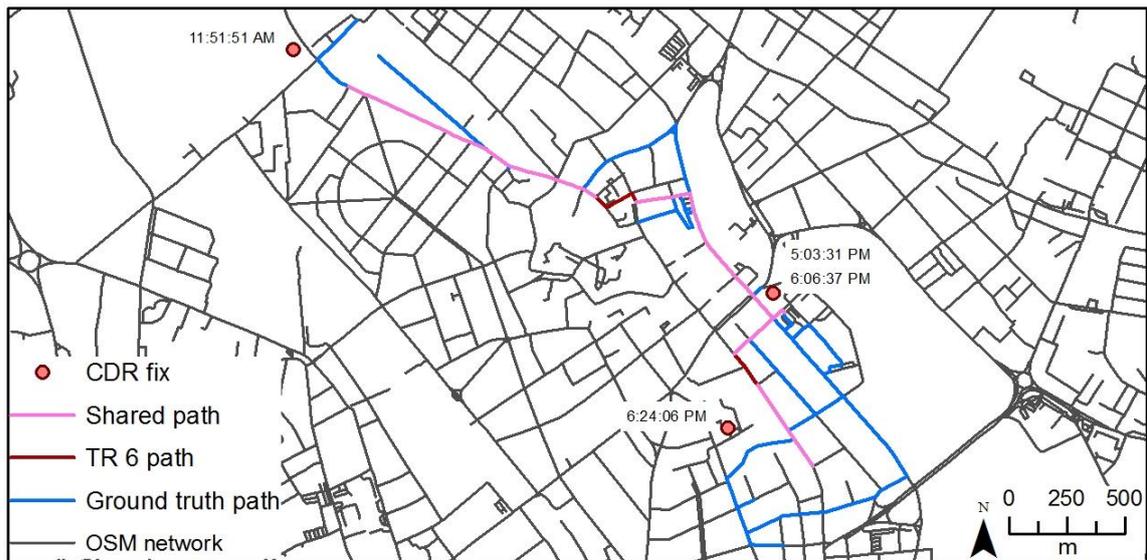
6.4.2 Impact of CDR data properties on accuracy of trajectory reconstruction

As seen in the previous section, the average similarity measures for the different trajectory reconstruction methods are always lower than 0.3. The overall accuracy of the trajectory reconstruction methods developed in this thesis, therefore, is not very high. In the following, the intention is to investigate whether CDR data conditions can be found which may be expected to lead to a higher accuracy of trajectory reconstruction. The properties investigated in this study are the number of CDR fixes from different locations (Section 6.4.2.1), the temporal resolution of the CDR data (Section 6.4.2.2), and the scale of the mobile phone user's movement (Section 6.4.2.3). Figure 27 - Figure 29 exemplarily show configurations of reconstructed and ground truth paths, the original CDR data, as well as all the computed SMs 4 and 12 and the associated statistical properties that are investigated. The reconstructed paths are all based on TR method 6. The edges that are shared between the reconstructed and the ground truth paths are depicted in pink. The reconstructed trajectory in Figure 27 with SMs 4 and 12 of 0.83 and 0.84, respectively, is based on 11 CDR fixes from different locations. The associated temporal resolution (avg. time between consecutive CDR fixes) is 34.4 min and the total distance covered is approx. 200 km (in terms of total distance between consecutive GPS fixes). Similarity measures for the reconstructed path in Figure 28 are significantly lower with SMs 4 and 12 of 0.25 and 0.38, respectively. The number of spatially unique CDR fixes with 3, as well as the temporal resolution with on average 130.8 min between consecutive calls, are considerably lower and the total distance covered of 9 km is much shorter in comparison to the statistical properties of Figure 27. The reconstructed path in Figure 29 has the lowest similarity measures of the three examples. The total distance covered is only 3 km and the number of unique CDR fixes is 2. The temporal resolution is slightly lower than that of Figure 27. The figures are used to illustrate the discussion of the results in the following three sections.



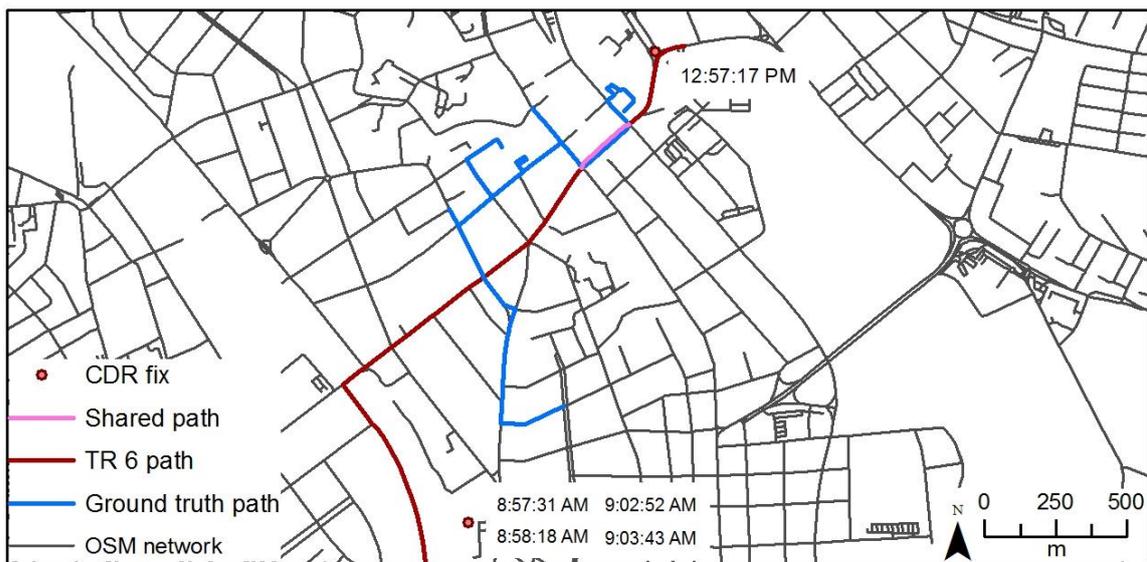
SM 4	SM 12	Unique CDR fixes	Temp. resol. [min]	Total distance [km]
0.83	0.84	11	34.4	203.8

Figure 27: Path reconstructed by TR method 6 and corresponding ground truth path, plus associated similarity measures and statistical properties of the data used (based on CDR and GPS data of user 4, 11.06.2013)



SM 4	SM 12	Unique CDR fixes	Temp. resol. [min]	Total distance [km]
0.25	0.38	3	130.8	9.2

Figure 28: Path reconstructed by TR method 6 and corresponding ground truth path, plus associated similarity measures and statistical properties of the data used (based on CDR and GPS data of user 6, 22.06.2013)



SM 4	SM 12	Unique CDR fixes	Temp. resol. [min]	Total distance [km]
0.07	0.32	2	49.0	3.0

Figure 29: Path reconstructed by TR method 6 and corresponding ground truth path, plus associated similarity measures and statistical properties of the data used (based on CDR and GPS data of user 5, 28.06.2013)

6.4.2.1 Impact of number of spatially unique CDR fixes on accuracy of trajectory reconstruction

In a manner similar to Newson and Krumm (2009), this section tests how the quality of the resulting paths is influenced by the amount of available data serving as input for the methods developed during this research. In their study, Newson and Krumm validated map-matching methods by comparing map-matched paths to ground truth paths, by actively degrading their data by reduction of the temporal resolution. By contrast, this study relies on the variety of the available number of fixes in the 73 daily segments under investigation.

Since multiple consecutive CDR fixes with identical locations serve as a single input information for the TR methods, the number of spatially unique CDR fixes might therefore be the more relevant indicator for the input data quality – in terms of quantity of data – than the total number of CDR fixes. This variable, however, is only a proxy for the number of CDR fixes that are actually used as input for the methods. In fact, the precise indicator for the amount of usable input data would be the number of consecutive phone activities from different locations. The measure used here (number of spatially unique CDR fixes), however, is probably a good proxy for the number of consecutive phone activities from different locations, since it can be assumed that phone activities from identical locations usually occur directly one after the other. In addition, although the movement back and forth between two locations during an entire day could be reconstructed, the similarity assessment (as indicated in Section 6.3.2) is purely spatial and therefore does not take into account multiple travelled edges.

The exemplary ground truth – reconstructed path comparisons in Figure 27 - Figure 29 may give rise to the assumption that the quality of trajectory reconstruction is improved when the number of spatially unique CDR fixes available is augmented. Figure 30 and Figure 31 investigate whether this assumption applies to all daily segments of the test sample. Figure 30 shows the average SMs 4 and 12 for an augmented exclusive criterion of minimum required number of CDR fixes from different locations for TR method 6. When assuming to tolerate daily segments with at least 2 spatially unique phone activities, which corresponds to the minimum condition for TR and represents all daily segments available, average SMs of 0.21 and 0.29 are obtained. An upward trend in the average SM with increase of the minimum required CDR fixes from different locations is clearly distinguishable. Daily segments, featuring 11 spatially unique CDR fixes, reach SMs 4 and 12 of 0.77 and 0.73, respectively. In terms of SM 4, this means that 77% of the ground truth paths could be reconstructed on the basis of the CDR fixes, which is quite promising. This finding must be treated with caution though, because the average SMs for 11 available CDR fixes with unique location is based on two daily segments only.

Figure 31 shows again the scatter plot for the total of 511 comparison cases, this time color-coded according the number of available spatially unique fixes, classified into groups of equal to or less than 4 (colored in red) and more than 4 (colored in blue) available spatially unique CDR fixes. The scatter plot clearly shows a cluster of blue points in the upper right part of the figure. Analysis of both Figure 30 and Figure 31 provides evidence that the introduction of a criterion of minimum required number of CDR fixes with unique locations per daily segment for TR leads to a considerably higher path quality.

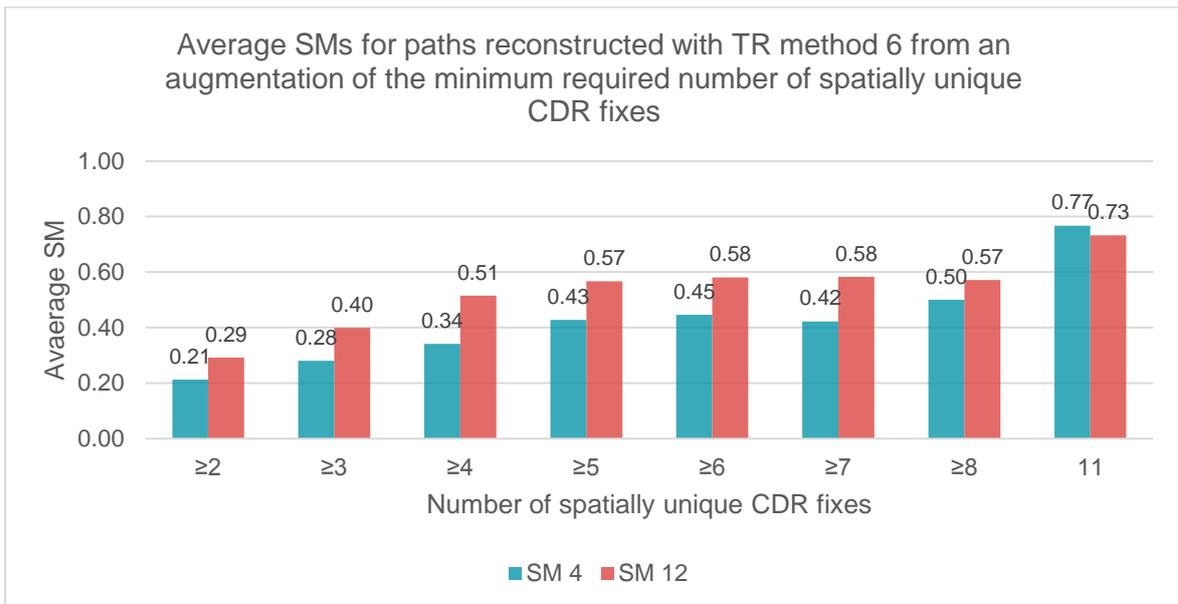


Figure 30: Bar chart for TR method 6 with SMs 4 and 12 resulting from an augmentation of the minimum number of required CDR fixes with unique locations, thereby reducing the number of daily segments considered

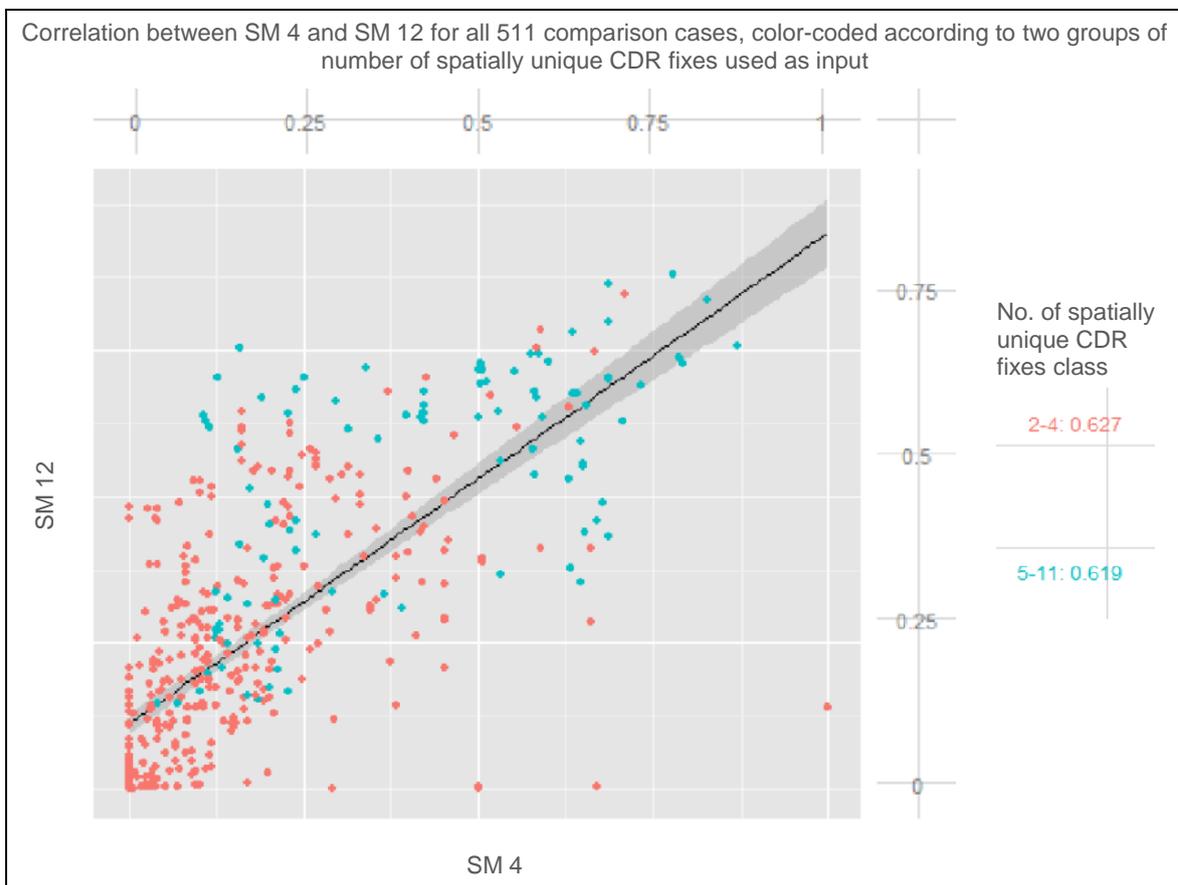


Figure 31: Scatter plot for SMs 4 and 12, color-coded according to two classes of number of spatially unique CDR fixes used as input

6.4.2.2 Impact of temporal resolution of CDR data on accuracy of trajectory reconstruction

A further variable that might give an indication on the quality of the CDR segments is the temporal resolution of the included fixes. The underlying assumption is that data available in smaller time intervals give more precise information about the movement behavior of the respective mobile phone users. The average time difference between consecutive CDR fixes of a daily segment is used as a proxy for the temporal resolution of the input data. The fact that consecutive phone activities with identical locations in a daily segment are not excluded from the computation of the temporal resolution brings some bias into this variable. This is due to the fact that multiple phone activities via the same antenna, despite high temporal resolution do not increase the level of information of input data (as stated in the previous Section 6.4.2.1). Since this applies equally to all of the daily segments, this shortcoming should not introduce significant interference.

To test whether the temporal resolution has an impact on the quality of the reconstructed paths, the daily segments were classified into three equal-sized groups of high, medium and low temporal resolution and subsequently analyzed. The respective class boundaries, given by the average time difference between consecutive phone activities, are indicated in Table 14. It may be observed that temporal resolutions vary dramatically between 6 min and more than 9 h. The average SMs 4 and 12 of the reconstructed paths with TR method 6 for each group are shown in Figure 32. It is notable that daily segments with medium temporal resolution of 55-110 min seem to produce the best reconstruction quality. It is counterintuitive that segments with a high temporal resolution have lower similarity values. A possible explanation could be that multiple calls in short temporal intervals are an indication that the phone user is stationary rather than mobile, busy calling people, and therefore is not giving any information of value regarding his movement behavior (in terms of different antennas used). The CDR fixes in Figure 29 show this type of calling pattern: 4 phone activities within 6 min were routed via the same antenna. This finding would be in contradiction to the statement from the previous paragraph that multiple phone activities via the same antenna are equally distributed throughout the data sample, but are much more prevalent in the group of high temporal resolution segments. In further research, it would be interesting to compute temporal resolution considering only consecutive CDR fixes with different locations, testing whether this has an impact on the ranking in terms of the average SMs of the three temporal resolution groups.

Figure 33 shows again the scatter plot for the paths reconstructed using all the seven different TR methods, this time color-coded according to the temporal resolution. A clear pattern is not distinguishable. The high (red) and medium (blue) resolution dots are slightly overrepresented in the upper right part, whereas the low resolution (green) dots are rather to be found in the lower left part. These are only tendencies and a clear statement about the impact of the temporal resolution of the CDR data on the reconstruction quality cannot be made. The visualized examples in Figure 27 (high temp. resolution, SMs of 0.83 and 0.84), Figure 29 (medium temp. resolution, SMs of 0.07 and 0.32), and Figure 28 (low temp. resolution, SMs of 0.25 and 0.38), neither give evidence that medium temporal resolution

yields the highest reconstruction quality, nor do they support the initial assumption that trajectory reconstruction is improved when the temporal resolution is augmented.

Table 14: Class boundaries for the three equal-sized groups of low, medium and high temporal resolution according to the average time difference between consecutive CDR fixes

Low temp. resolution	Medium temp. resolution	High temp. resolution
580-111 min	110-55 min	54-6 min

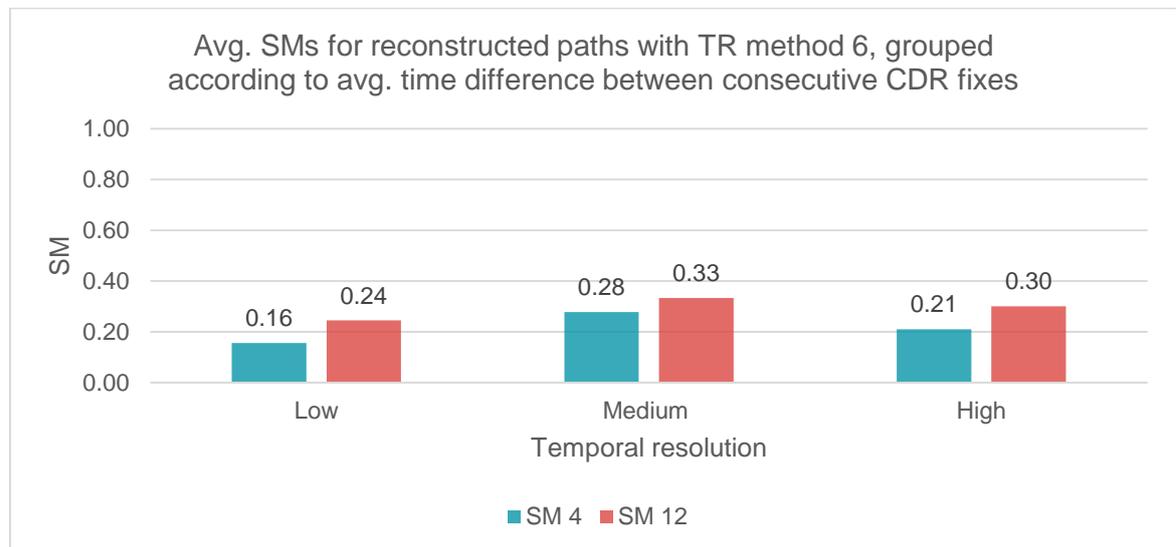


Figure 32: Bar chart representing the avg. SMs for reconstructed paths with TR method 6 for three groups of avg. time between consecutive CDR fixes

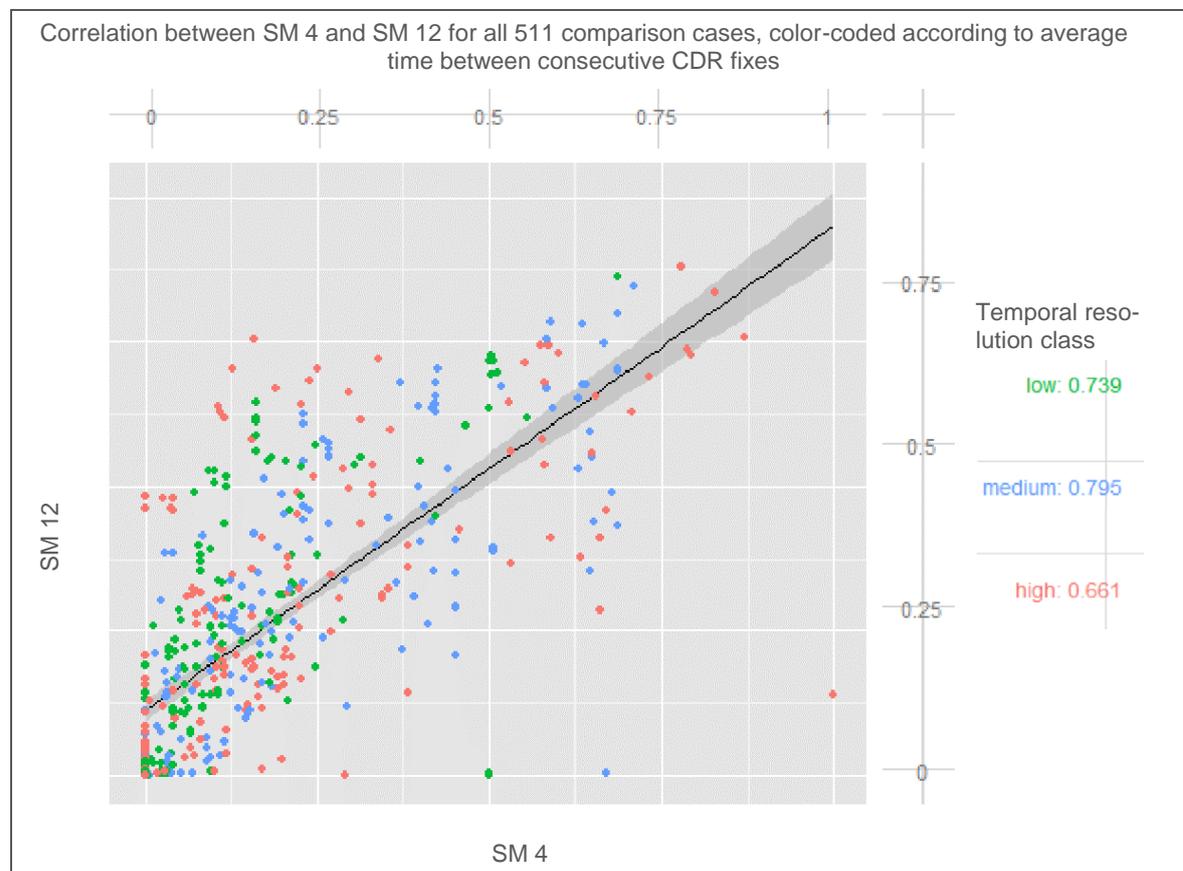


Figure 33: Scatter plot for SMs 4 and 12 of all 511 comparison cases, color-coded according to three classes of average time between consecutive CDR fixes

6.4.2.3 Impact of scale of movement on accuracy of trajectory reconstruction

Besides the impact of the characteristics of the CDR data on the quality of the reconstructed paths, which was investigated in the previous Sections 6.4.2.1 and 6.4.2.2, the nature of the movement itself might have an impact on the quality of the result. The expressed movements are especially heterogeneous in terms of their geographical extent or the size of their spatial footprint. The scale of a trajectory can be assessed, inter alia, in several ways: the total area covered (e.g., described by the minimum bounding box, the concave or the convex hull of the CDR or the GPS fixes), the total length of the CDR or the GPS paths, or the sum of the Euclidean distances between consecutive CDR or GPS fixes. In this study, the sum of the Euclidean distances between consecutive GPS fixes, which is an approximation for the total length of the movement expressed in reality, was used to categorize the daily segments into three groups of equal size (ca. 24 segments per group). As listed in Table 15, short-distance, medium-distance, and long-distance movements entail path lengths derived from the GPS fixes of 0.7-5.5 km, 5.6-14.9 km, and 15.0-482.0 km. Short-distance movements would typically represent intra-city movements. The medium-distance group comprises typical lengths of commuting distances from suburban to urban areas. And the third category would represent inter-city trips.

Figure 27 - Figure 29 seem to show a clear trend of an improved trajectory reconstruction quality when the map scale on which the movement is represented gets smaller, and the distance covered increases. Figure 34 and Figure 35 investigate whether this assumption can be affirmed when all daily segments are considered. The bar chart in Figure 34 shows that the average SMs 4 and 12 for the paths reconstructed with TR method 6 increase considerably when the distance of the movement is varied from short to long. As indicated by SM 12, the separate convex hulls of the reconstructed and the ground truth paths belonging to the long-distance group, on average have 41% of overlap, whereas the paths belonging to the short-distance group only have 21% of overlap on average. The scatter plot in Figure 35 represents again the values of SMs 4 and 12 for all 511 comparison cases. A cluster of red dots (representing long-distance movements) is clearly identifiable in the upper right part of the diagram. The blue dots (representing short-distance movements) and the green dots (representing medium-distance movements) are mainly to be found in the lower left area of the diagram. Both Figure 34 and Figure 35 give evidence that there is a positive correlation between the geographical dimension of the movement and the quality of the reconstructed paths.

Table 15: Class boundaries for the three equal-sized distance groups according to the path lengths derived from the GPS fixes

Short-distance	Medium-distance	Long-distance
0.7-5.5 km	5.6-14.9 km	15.0-482.0 km

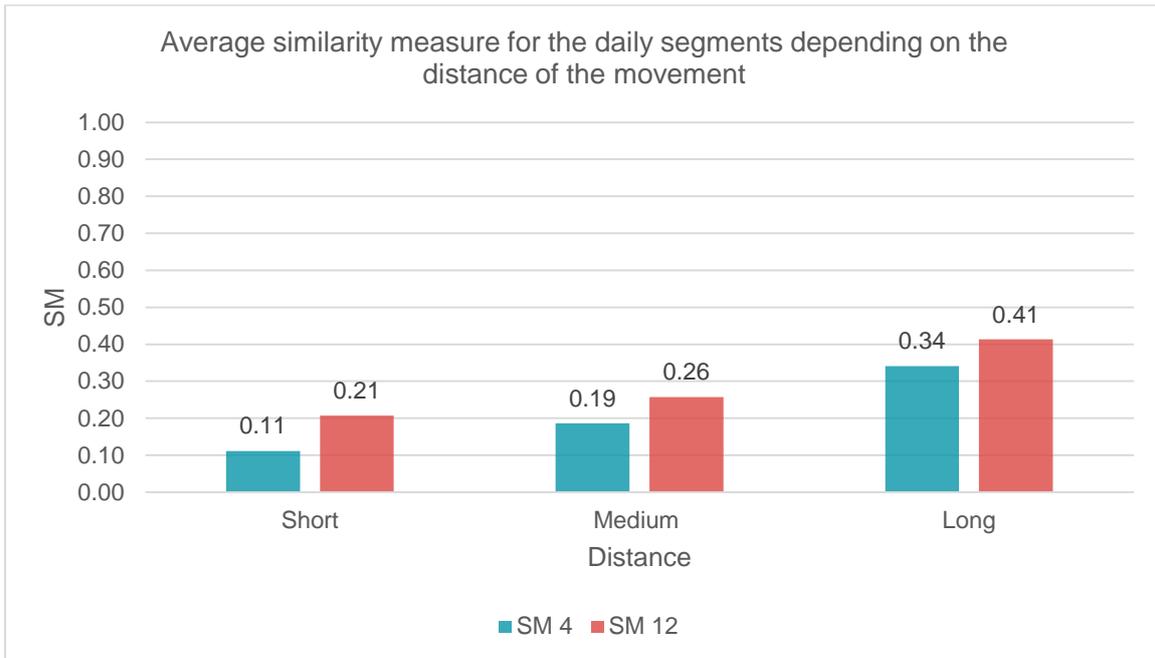


Figure 34: Bar chart with average SMs 4 and 12 for daily segments grouped into three classes of distance of movement

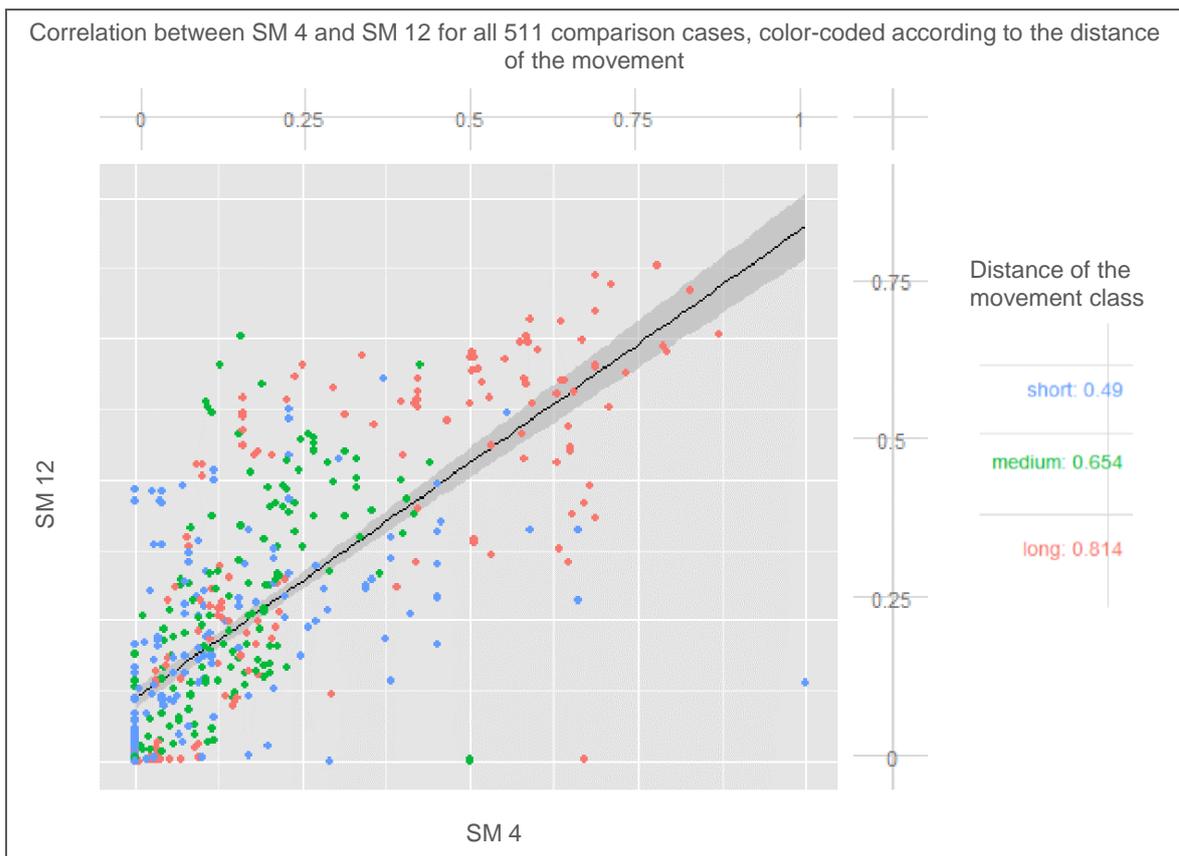


Figure 35: Scatter plot representing the correlation between SMs 4 and 12 for all 511 comparison cases, color-coded according to three distance of movement classes

7 Discussion

This chapter addresses the research questions (RQs), placing the results of the study in the context of the related research. Firstly, the trajectory reconstruction methods devised in this study are discussed and suggestions for improvements are made (Section 7.1). Subsequently, Section 7.2 discusses the validation of the trajectory reconstruction methods and gives suggestions for further investigation. Finally, Section 7.3 discusses whether CDR data properties have an impact on the quality of the reconstructed trajectories and directions for future studies are identified.

7.1 Methods to reconstruct trajectories from sparse CDR data

Reconstructing trajectories from CDR data which are sampled in irregular temporal intervals and with a spatial resolution equal to the size of the coverage area of an antenna requires several pre-processing steps (as described in Chapter 4) and application of a number of heuristics (as described in Chapter 5). In contrast to the statement of Blumenstock (2012) that there is no knowledge available regarding the whereabouts of a user between phone activities, in this study several assumptions are made in order to narrow down the user's potential location between consecutive phone activities. Such additional knowledge is derived, for example, from the assumptions that individuals' movements are usually bound to a network (Brinkhoff (2002) or Jiang and Jia (2009)) or that individuals tend to follow the shortest path when travelling between two locations. These and further assumptions serve as basis for the discussion of RQ 1.

RQ 1: How can mobile phone users' trajectories be reconstructed from sparsely sampled CDR data?

Assumption of network-bound movement

In order to narrow down the potential whereabouts of a mobile phone user, it is assumed in this study that movements take place on a network. Therefore, the OSM road network is applied to which the CDR locations are projected and on which the shortest paths between the identified nodes are computed. The assumption of network-bound movement is certainly a reasonable one since it is observable in everyday life and also confirmed from literature (e.g., Brinkhoff, 2002). A visual inspection of the GPS fixes supports the assumption that the trajectories of our test users follow the road network in most of the cases. One should be aware, however, that errors are introduced by movements that are not bound to a network. Additionally, there are methodological constraints regarding the network used in this thesis. That is to say that the network basically consists of roads intended for cars which constitutes a problem for movements expressed by pedestrians or train users. Broadening of the network to include a wider range of travel possibilities would certainly be part of further investigations

Segmentation of the CDR data

A first important decision to enable trajectory reconstruction from CDR data, which needs to be made, concerns the choice of an appropriate temporal frame of reference. This study

used a segmenting approach based on daily segments. Therefore, midnight, which is supposed to be a time when people typically sleep in their homes, was chosen as segmenting criterion. Consequently, trajectories are reconstructed that start in the morning when an individual usually begins to move by leaving his home and end in the evening when the individual typically is home again. As a minimum condition for the trajectory reconstruction, at least two mobile phone activities routed via different antennas need to be available in the daily segments. Otherwise, it is not reasonable that the user was present at any locations other than the one covered by the single antenna. A limitation of the segmenting criterion used in this study is that, particularly at weekends, when people tend to stay awake and move around longer, midnight is not an appropriate time to divide the trajectories. Another segmenting approach could be the estimation of the home of a user by analyzing over which antenna the majority of phone activities are routed outside usual office hours, as in Ahas et al. (2009) or Csáji et al. (2013). The segmentation could subsequently be based on the spatial criterion of the “home antenna”. González et al. (2008) find that human trajectories show high temporal and spatial regularities, therefore it would be interesting to try to aggregate CDR data, for example, of multiple workdays or of multiple Mondays. The application of alternative temporal frames of reference is certainly an interesting direction for future research.

Heuristic-based map matching to deal with coarse spatial resolution

To deal with the coarse spatial resolution of the CDR data, different map-matching approaches are employed in order to locate the user in a realistic position on the road network as suggested by Saluveer and Ahas (2014). As stated by Zang et al. (2010), there is very little research on the localization of a user solely on the basis of the antenna position obtained from CDR log files. Most researchers use additional information such as RSSI. One achievement of this thesis is the proposition of a number of map-matching heuristics for CDR data and their validation with GPS data. The map-matching algorithms which were proposed are based on the use of the Voronoi cells which are quite frequently employed to describe the area which is closest to an antenna compared to others and therefore the area with the highest probability for the location of the mobile phone user. Voronoi cells, however, are an ideal approximation and therefore may be unreliable indicators of the actual area in which the user is located. Especially in urban areas with a dense cellular network, it is possible that the mobile phone switches to an antenna that is actually not the closest to the mobile phone, if the line of the antenna signal is interrupted (through noise introduced by scattering and reflection of buildings) or if the closest antenna is very crowded (Blumenstock, 2012; Csáji et al., 2013). In this case, it is obvious that the map-matching methods used in this study are incapable of finding the true position.

A set of candidate nodes / edges is defined for each CDR fix by intersecting the corresponding Voronoi cell with the nodes / edges of the road network, respectively. A number of different criteria are used to rate a node / edge of the set of candidate nodes / edges, respectively, as the most probable location of the phone user. A criterion used for nodes is, for example, the degree centrality. According to Crucitti et al. (2006), this centrality measure is an inappropriate measure in urban networks because node degrees are limited due to geographical constraints. It would be interesting to investigate map-matching algorithms

based on further centrality measures proposed by Crucitti et al. (2006), e.g., closeness centrality. Edges were favored according to their attributes. As suggested by Saluveer and Ahas (2014), in the edge-based map-matching methods semantically more important roads (MM method 5) or roads that permit a higher speed are favored (MM method 6). The underlying assumption is that more important roads are more frequented and the probability of an accurate positioning, therefore, is higher.

The pre-validation of the map-matching methods (Section 5.2.6) show that most of the matched CDR locations are within a reasonable distance from the temporally closest GPS fixes. The mean and median distances between the map-matched CDR fixes (with the best performing MM method 6) and the ground truth of 938.5 and 412.7 m, respectively, indicate that the map matching leads to an improvement of the locational accuracy when comparing these numbers to the length of a cross-section of a cell size reported to be between 200 and 10'000 m by Saluveer and Ahas (2014). In further research, it would be interesting to similarly compute the distances between the ground truth and CDR fixes that are randomly assigned to a position within the Voronoi cell in order to verify whether the heuristic-based map-matching methods perform better.

Shortest-path heuristic to fill temporal gaps between consecutive phone activities

In order to deal with the low temporal resolution of the data, which results in large spatial gaps between the known locations, a shortest-path algorithm is used to model the user's trajectory within two consecutive phone activities. Dijkstra shortest-path algorithm which is implemented in GeoTools (2014) – an Open Source Java Library – was therefore applied. Dijkstra shortest-path algorithm computes the shortest paths from one source location to all destinations on the road network and therefore needs to be reinitialized for every new source location. The algorithm implemented in GeoTools needs to be adapted in order to compute the shortest path between an ordered set of fixes. Additionally, a rule needs to be implemented that ensures that consecutive fixes with identical locations are only considered once, since a shortest-path computation between two similar locations is not sensible. The edges of the road network are weighted according to their length. If maximum speed tolerated was known for all edges, it would be sensible to use average travel time to traverse an edge instead of the length. In this way the fastest instead of the shortest path would be computed. If more data were to be processed and computational efficiency consequently was more important, the application of a faster algorithm such as A* algorithm or the algorithm proposed by Geisberger (2008) would be recommended. Instead of the shortest-path heuristic, which is also used by Järv et al. (2012) to predict the movement between home and work location of Tallinn commuters, a “simplest path” heuristic as proposed by Duckham and Kulik (2003) could be implemented in future research.

Summary and Outlook

As shown in this thesis, trajectory reconstruction from sparsely sampled CDR data is feasible by firstly, map matching CDR fixes based on various assumptions to the road network and in a second step, connecting the map-matched locations with a shortest-path algorithm. The few other trajectory reconstruction algorithms proposed in literature so far are mostly designed for a particular purpose (e.g., inferring travel mode between a pre-defined origin and destination pair in Doyle et al. (2011) or Wang et al. (2010)) and therefore are only

applicable under constrained conditions. In a further development of the trajectory reconstruction methods, consideration should be given to how the temporal information of the CDR data could be maintained in the trajectory. The edges composing the paths could be assigned time information by interpolation between the last phone activity of the previous antenna and the first phone activity of the following antenna. However, such interpolation would require most uncertain and speculative assumptions. If, for example, a temporal gap of several hours occurs between two consecutive calls taking place at a distance of 10 km, the assumption of linear movement behavior between the two consecutive phone activities as used in the decompression algorithm for previously compressed GPS data in Richter et al. (2012) would probably not be a sensible one in this case. Maybe the inclusion of the behavior of a user during previous days would help to gain knowledge regarding how much time is usually spent in a particular place. Another methodological issue is how stop-overs (inferable from multiple phone activities routed via the same antenna over a certain time period) could be modeled.

7.2 Validation of the trajectory reconstruction methods

In the previous section, it was discussed how trajectories can be reconstructed from CDR data. The logical question that follows is how well the devised methods perform. With the exception of a few studies (e.g., Zang et al., 2010), no validation of methods developed in the context of transportation research for mobile phone data has been carried out (Smoreda et al., 2013). The availability of GPS data for the six test users, generally with a much higher spatial and temporal resolution, enables a validation of the methods developed in this research by assessing the similarity between reconstructed trajectories and the ground truth. This contributes to the discussion of RQ 2:

RQ 2: In order to validate the trajectory reconstruction methods developed in this study, what level of similarity can be achieved by comparison of the reconstructed trajectories with higher resolution GPS trajectories of the same journeys?

Before discussing RQ 2, it is reasonable to consider the following issues:

- How suitable are the used comparison units?
- How do the SMs assess the similarity?

Suitability of comparison units

In order to facilitate the comparison of the ground truth to the reconstructed trajectories, the GPS points are transformed into a structure similar to that of the reconstructed trajectories, which is a subset of connected edges from the OSM road network (see Section 6.2). Therefore, the GPS fixes are matched with the road network using a geometrical map-matching algorithm that identifies the set of road network edges that are closest to at least one GPS fix. Based on an edge-score criterion (number of GPS points projected to a specific edge in relation to its length) unlikely edges are disqualified from the set. In a final step, the spatial gaps in the ground truth path are filled by application of a shortest-path heuristic. A major constraint of the map-matching methodology developed in this study is that multiple travelled paths are not identifiable. The quality control of the ground truth path by comparison

to the original GPS data (cf. Section 6.2.4) shows that the average distance of the GPS paths to the original data is at maximum 54 m. The approach could certainly be improved, for example, by taking into account edge connectivity, as is suggested by Velaga et al. (2009). Another possibility would have been to not map match the GPS points to the network and thereby avoid manipulation of the ground truth data. However, similarity comparisons such as path alignment would not have been possible using this approach. Similarity could be assessed by computing the average minimum distance from a GPS point to the reconstructed path. On the basis of such a measure, it would be difficult to say how similar two paths really are, and such an approach would be very sensitive to GPS sampling issues.

Effectiveness of similarity assessment

As seen from the review of literature regarding the assessment of trajectory similarity in Section 2.4, numerous measures to compare trajectories are used. In this study, a range of similarity measures with their associated advantages and disadvantages (see Table 11, Section 6.3.2) have been proposed and implemented. If the order of the ground truth path could be maintained, it would be interesting to implement Fréchet distance instead of Hausdorff distance which is said to better assess similarity between two planar curves as observed by visual inspection (Alt et al., 2004). Both measures, however, indicate similarities between two paths as distances in a metric unit such as km or m. In order to receive values between 0 and 1 for low and high similarities, respectively, a normalization of the resulting distances would be required. Since no obvious way to normalize the resulting distances exists so far, Hausdorff and Fréchet distances make comparisons between trajectories of very different nature (e.g., in terms of length), as it is the case in the data sample, and comparisons to other similarity measures difficult. A further interesting idea would be to include temporal information in the similarity assessment by comparing the reconstructed path segments to the ground truth path segments for time intervals defined by two consecutive CDR fixes. A more in-depth discussion of the various proposed similarity measures is to be found in Section 6.3.3.

SM 4 computes the ratio of the length of the shared edges between the ground truth path and the reconstructed path to the total length of the ground truth path. It was chosen as one of the SMs on which further analyses are based on, since it weighs edges according to their length and uses the ground truth path as reference that should be approximated as closely as possible. This SM is comparable to the measure used in Lou et al. (2009) to quantify the quality of their map-matching algorithm. The only difference is that the length of correctly identified edges is divided by the length of the “reconstructed” path and not by the length of the ground truth path. SM 4 is a reliable similarity measure in that a high value of the SM always indicates that most of the edges could be precisely reconstructed on the basis of the CDR data. It is a very strict similarity measure though and as soon as the edges are not identical, it cannot assess whether two trajectories are still close to each other or very far apart. For this reason, SM 12 (ratio of the area of intersection of the convex hull of the ground truth path and the convex hull of the reconstructed path to the area of the union of the two separate convex hulls) was chosen as the second SM on which to base the validation and subsequent analyses of the impact of different CDR data properties on the trajectory reconstruction quality. SM 12 compares the areas of the movement of two trajectories to each other rather than comparing exact path alignment and is therefore more tolerant in

most cases. Figure 29 (see Section 6.4.2) shows nicely how SM 4, with a value of 0.07, exhibits a much lower similarity for the same ground truth and reconstructed paths constellation than SM 12 with a value of 0.32. Depending on the type of similarity required (which again depends on the purpose of the similarity assessment and the type of trajectory), SM 4, SM 12 or even a very different one might be the most suitable.

Validation of trajectory reconstruction methods

The quality of the trajectory reconstruction methods is assessed by computation of the SMs 4 and 12 for the reconstructed and the corresponding ground truth paths. As found in Section 6.4.1, the average similarity measures for the different trajectory reconstruction methods are 0.15-0.21 and 0.23-0.29 for SM 4 and SM 12, respectively. Even TR method 6 (based on the fastest edge heuristic) which was qualified as the method yielding the highest similarity measures, has values of SM 4 and 12 lower than 0.30 and 0.44, respectively, for 75% of the cases. These rather low values indicate that the methods developed in this study do not work well on a general level. The low similarity measures may be issued from an averaging of the similarity measures over the total number of the comparison cases. The scatter plot in Figure 25 shows that in some cases high similarity measures are yielded. In the following Section 7.3, it is discussed whether CDR data conditions could be established, under which higher similarity measures are to be expected.

Summary and Outlook

In conclusion, the similarity measures to be expected from the TR methods proposed in this study are, in general, rather low. However, it must be acknowledged that the ground truth paths are only approximations of the actual travelled paths and that the SMs are two possible ones amongst many others that could capture similarity in very different ways. As discussed in the previous section, TR methods could be improved using improved map-matching and gap-filling heuristics, but the low temporal and spatial resolution of the CDR data impose strict limits on the potential trajectory reconstruction quality that can be expected. If an individual does not use his phone while visiting a certain place, there is no means to extract that location from the CDR data if the place is not located somewhere within two available CDR fixes. The few researchers who comparably reconstructed trajectories on the basis of CDR data (e.g., Blumenstock, 2012; Csáji et al., 2013; González and Barabási, 2007), notably under constrained conditions (cf. Section 2.2.2), did not validate their methods. It would be interesting to apply the TR methods developed in this study to different mobile phone users and to other road network settings (from different countries) in order to compare trajectory reconstruction qualities to the ones obtained here.

7.3 Impact of CDR data properties on trajectory reconstruction accuracy

In Section 6.4.2, the reconstructed trajectories were systematically grouped according to different CDR data properties, in order to investigate whether these properties have an impact on the trajectory reconstruction quality. This leads to the discussion of the final research question:

RQ 3: Which properties of the CDR data, such as sampling properties or trajectory length, affect the accuracy of the reconstructed trajectories?

Trajectories obtained from CDR data are very heterogeneous as they result from different users' backgrounds (movement behavior, etc.) and different mobile phone use behavior (number of CDR fixes, etc.). In order to test whether certain kinds of CDR data facilitate more accurate trajectory reconstruction, the following three properties of the CDR segments were examined: number of spatially unique CDR fixes, temporal resolution of CDR data, and scale of the movement. These are discussed in the following:

Number of spatially unique CDR fixes

In Section 6.4.2.1 it was ascertained that the number of spatially unique CDR fixes has a considerable impact on the obtained similarity measures. With at least 5 CDR fixes from different locations, for TR method 6, similarity measures around 0.5 are to be expected, instead of 0.2-0.3 when accepting all daily segments (at least 2 fixes). By raising the number of spatially unique CDR fixes to 11, the accuracy could be improved to approx. 0.7. This outcome is mostly in line with the findings of Saluveer and Ahas (2014) who state that at least 15 CDR fixes per day are required to adequately reconstruct a user's trajectory. According to their study, trajectory reconstruction with fewer than 7 fixes becomes problematic. Multiple phone activities via the same antenna are included in these numbers. This could provide a partial explanation for why higher threshold values are proposed by Saluveer and Ahas compared to those in this study. In their study, however, no specific indications are to be found for how trajectories were reconstructed and validated.

Newson and Krumm (2009) emphasize the importance of assessing when a method breaks down by giving an indication of the minimum temporal resolution required for reasonable results. In their transportation mode inference study, Wang et al. (2010), for example, include only users who engage in at least one phone activity per hour in order to have more spatio-temporal information that enables to more accurately infer the trips the users made. Since they have access to a dataset of close to one million users, they can afford such a strict filtering criterion. A further way of finding such a minimum number of required CDR fixes for the trajectory reconstruction methods proposed here could consist of using data detail records (DDR). These data have the same spatial resolution as CDR data, but typically feature a much higher temporal resolution, since many services on the mobile phone regularly connect to the internet. On average, 100 DDR fixes in contrast to 6 CDR fixes are registered per person per day (Saluveer and Ahas, 2014). By reconstructing trajectories from DDR data that are iteratively reduced in the number of available fixes, it is possible to define the threshold number of fixes from which a reasonable or expected accuracy can be obtained.

Average time between consecutive CDR fixes

The grouping of the CDR segments according to their temporal resolution showed fewer important effects on the average similarity measures obtained than the number of CDR fixes (see Section 6.4.2.2). The assumption that an increase of the temporal resolution would improve the reconstruction quality could not be affirmed. Instead, it was found that a medium temporal resolution (avg. of 55-110 min between consecutive phone activities) yields

the best reconstruction quality, with SMs still very low around 0.3 for TR method 6. A proposed explanation for the lower reconstruction quality with higher temporal resolution data was that phone activities taking place in short time intervals typically originate from the same place and therefore do not give indications of the mobility behavior of a user. The finding that the temporal resolution does not significantly affect the accuracy of the reconstructed trajectories could result from the fact that the comparison of the paths is purely spatial and time is not taken into account.

Scale of the movement

The last criterion that was investigated was the effect of the scale of the movement on the trajectory reconstruction quality. As proxy for the scale of the movement, the distance between consecutive GPS points is summed, in a manner identical to the measure “*total line segment length*” as proxy for the distance covered by an individual described in Csáji et al (2013). The findings in Section 6.4.2.3 show that the group of long-distance movements (referring to inter-city trips) yields a considerably higher trajectory reconstruction accuracy than the group of medium-distance and short-distance movements (referring to commuting traffic and intra-city trips, respectively). These findings are in line with the ones of many researchers (e.g., Rose (2006) or Zang et al. (2010)) who likewise ascertain that trajectory reconstruction for inter-city trips is easier than for inner-city trips. Possible factors explaining this observation that are mentioned by the authors are the significantly higher road network densities and the higher likelihood of a mobile phone not connecting to the closest antenna due to overcrowding or disruptions of the antenna signals caused by buildings. Both factors apply to this study as well and constitute a difficulty for a correct map matching of the CDR fixes to the road network. Additionally, it is conceivable that long-distance travelers in the majority of the cases follow the roads that primarily permit fast movement and therefore the map-matching methods mainly relying on the importance of a node or edge (e.g., highest road category, highest speed tolerated, etc.) function better. In contrast, short-distance movements (intra-city trips) take frequently place on less important roads as well, since the primary goal is to get to the destination which is not necessarily next to a major road. An implication of this finding might be that trajectory reconstruction from CDR data is generally better suited for long-distance movements.

Summary and Outlook

In conclusion, the findings show that better trajectory reconstruction accuracies are to be expected if more spatially unique CDR fixes serve as input for the methods and when the movement takes place over a long distance. In order to verify these findings, the examination of more CDR segments would be required. In further research it would be interesting to test whether the data properties investigated in this study correlate amongst each other (e.g., relationship between long-distance movements and number of phone activities). Furthermore, it would be interesting to examine further criteria such as the impact of urban vs. rural settings on the accuracy of short-distance movements.

8 Conclusion

8.1 Summary

There have been many attempts to reconstruct people's movements on the basis of positioning data, relying on positioning data such as GPS with high temporal and spatial resolution. Call detail records (CDRs), which are used in this study, are automatically stored for billing purposes by mobile phone operators and are therefore potentially available relatively cheaply and for long time periods for a very large fraction of the population (Furletti et al., 2012). This kind of data, however, is generally not easily obtainable from mobile phone operators and the use of it has important implications for the privacy rights of the people concerned. In this study, the CDR and the corresponding GPS data of 6 mobile phone users in Estonia over a one month period are provided by Positium LBS (2014) in collaboration with its long-standing partner in academia, the Geography Department of the University of Tartu and consent to use the data was obtained from the mobile phone users involved in this study. The major constraints of this data source are the low spatial resolution, which depends on the coverage area of an antenna and the antenna network density, and the generally low temporal granularity depending on the regularity/irregularity and the frequency of phone activities of a user.

It is therefore a challenge to gain knowledge of the movement behavior of an individual from the CDR data. In this thesis, several methods for reconstructing trajectories from sparse CDR data have been proposed and validated. To this end, the following steps have been undertaken: The one month CDR and GPS data of the 6 test users are divided into daily segments and subsequently clipped according to each other's time frames. CDR segments consisting of fewer than two fixes with unique locations, as well as GPS segments with large spatial gaps, are disqualified from further analysis. The trajectory reconstruction consists of a two-level approach. Firstly, one of the seven different map-matching techniques that are proposed is used to match the CDR fixes to the most reasonable nodes on the road network. Secondly, the identified nodes are consecutively connected with a shortest-path heuristic. In order to validate the proposed trajectory reconstruction (TR) methods, the reconstructed paths are compared to the corresponding GPS trajectories. This is done by computing a set of similarity measures. Based on two selected similarity measures – relying on the number of shared path edges and the shared area of movement of the two paths, respectively – analyses are carried out. Thereby TR method 6, which favors edges with higher speed limitations, is identified as the most satisfactory method. Furthermore, it could be found that an increasing number of spatially unique CDR fixes and movements of an increasing distance have a positive impact on the accuracy of trajectory reconstruction, whereas the temporal resolution of the CDR data is less important.

8.2 Contributions

This is one of the first attempts to propose and validate concrete and generally applicable methods to reconstruct trajectories on the basis of a set of pre-processed CDR segments and the inclusion of a road network. To this end, a combination of already available GIS

methods, firstly, to match CDR data to the road network, and secondly, a shortest-path algorithm to connect the identified nodes, is applied. Profiting from the special situation of the availability of the higher resolution GPS data of the same journeys, the TR methods are validated by comparing the reconstructed trajectories to the corresponding GPS trajectories. This is a response to the urgent need for validation of methods relying on CDR data as ascertained by Smoreda et al. (2013). The validation of the methods shows that on a general level similarity measures are not expected to be particularly high. This is unsurprising, since the CDR segments are very heterogeneous, being dependent on the movement behavior and the calling habits of a mobile phone user. From the testing of the effect of CDR data properties on the accuracy of the TR methods, it is possible, however, to give indications of the data conditions, such as number of spatially unique CDR fixes and scale of the movement, under which higher accuracies of trajectory reconstruction are to be expected. The application of the methods developed in this study to data details records (DDR, mobile internet data) that are stored by many mobile phone operators would be easily feasible as the spatial resolution is identical to that of the CDR data and seems to be very promising, since the number of registered fixes is considerably higher.

8.3 Outlook

In order to further validate the TR methods and to establish with more certainty the criteria under which TR methods work reliably, testing on CDR or DDR data of greater numbers of users, over longer time periods, and also in road network settings from different countries is required. Further investigation could determine whether application of more refined map-matching methods or a different heuristic – e.g., simplest path as proposed by Duckham and Kulik (2003) instead of shortest path – to connect the nodes would produce higher similarity measures for TR methods. Additional clues such as time budget and speed limitations could be used in order to reduce the potential area of a mobile phone user's whereabouts on a road network, comparable to the approach used by Kuijpers et al. (2010). An area for further examination could be an investigation of whether a different approach of segmenting CDR data would lead to more reasonable trajectory units. The data could be segmented based on a spatial criterion, for example, the antenna which is expected to be the user's home location. Further research could be directed towards the inclusion of the temporal dimension to the trajectories. Edges comprising the trajectory could be assigned with temporal indications that have been reasonably interpolated between known CDR fixes. Besides the suggested improvements of the TR methods developed in this study, an application of the methods for many different purposes is also conceivable. It would be interesting, for example, to investigate the degree of similarity of trajectories for multiple daily segments of the same users, in order to test the assumption of a high spatio-temporal regularity of human trajectories as ascertained by González et al. (2008). In order to analyze the frequency of use of different road network edges, the reconstructed paths of multiple users (e.g., originating from a similar area or having a common destination, or of all users who made at least two calls from different places) could be aggregated and suitably visualized. With the information about the socio-economic background (e.g., gender, language, age) of mobile phone users, it could be investigated whether differences between different socio-economic groups regarding movement behavior (e.g., trajectory length, area of movement) are distinguishable.

9 References

- Agrawal, R., Faloutsos, C., Swami, A., 1993. Efficient Similarity Search in Sequence Databases. In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 69–84.
- Ahas, R., Aasa, a., Silm, S., Aunap, R., Kalle, H., Mark, Ü., 2007. Mobile Positioning in Space–Time Behaviour Studies: Social Positioning Method Experiments in Estonia. *Cartography and Geographic Information Science*, 34(4), 259–273.
- Ahas, R., Aasa, A., Mark, Ü., Pae, T., Kull, A., 2007a. Seasonal Tourism Spaces in Estonia: Case study with Mobile Positioning Data. *Tourism Management*, 28(3), 898–910.
- Ahas, R., Aasa, A., Roose, A., Mark, Ü., Silm, S., 2008a. Evaluating Passive Mobile Positioning Data for Tourism Surveys: An Estonian Case Study. *Tourism Management*, 29(3), 469–486.
- Ahas, R., Aasa, A., Silm, S., Tiru, M., 2010a. Daily Rhythms of Suburban Commuters' Movements in the Tallinn Metropolitan Area: Case Study with Mobile Positioning Data. *Transportation Research Part C: Emerging Technologies*, 18(1), 45–54.
- Ahas, R., Laineste, J., 2006. Technical and Methodological Aspects of Using Mobile Positioning in Geographical Studies. In: Pae, K., Ahas, R., Mark, Ü. (Eds.), *Joint Space. Open Source on Mobile Positioning and Urban Studies*. Positium, Tallinn, Estonia, 37–43.
- Ahas, R., Laineste, J., Aasa, A., Mark, Ü., 2007b. The Spatial Accuracy of Mobile Positioning: Some experiences with Geographical Studies in Estonia. In: Gartner, G., Cartwright, W., Peterson, M.P. (Eds.), *Location Based Services and TeleCartography*. Springer Berlin Heidelberg, 445–460.
- Ahas, R., Saluveer, E., Tiru, M., Silm, S., 2008b. Mobile Positioning Based Tourism Monitoring System: Positium Barometer. *Information and Communication Technologies in Tourism*, 475–485.
- Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M., 2010b. Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1), 3–27.
- Ahas, R., Silm, S., Saluveer, E., Järv, O., 2009. Modelling Home and Work Locations of Populations using Passive Mobile Positioning Data. In: *Location Based Services and TeleCartography II*. Springer Berlin Heidelberg, 301–315.
- Alt, H., 2009. *The Computational Geometry of Comparing Shapes*. In: *Efficient Algorithms*. Springer Berlin Heidelberg, 235–248.
- Alt, H., Godau, M., 1995. Computing the Fréchet Distance between Two Polygonal Curves. *International Journal of Computational Geometry and Applications*, 5(1&2), 75–91.
- Alt, H., Guibas, L.J., 1996. Discrete Geometric Shapes: Matching, Interpolation, and Approximation - A Survey. In: Sack, J.-R., Urrutia, J. (Eds.), *Handbook of Computational Geometry*. Amsterdam, Netherlands, 121–153.

- Alt, H., Knauer, C., Wenk, C., 2004. Comparison of Distance Measures for Planar Curves. *Agorithmica*, 38, 45–58.
- Andres, K., Ahas, R., Tiru, M., 2009. Analysing repeat Visitation on Country Level with Passive Mobile Positioning Method: An Estonian Case Study, 140–155.
- Andrienko, G., Andrienko, N., Bak, P., Bremm, S., Keim, D., von Landesberger, T., Pölitz, C., Schreck, T., 2010. A Framework for using Self-Organising Maps to analyse Spatio-Temporal Patterns, exemplified by Analysis of Mobile Phone Usage. *Journal of Location Based Services*, 4(3&4), 200–221.
- ArcGIS, 2014. ArcGIS. Retrieved from: www.arcgis.com (accessed 8.5.2014).
- Asakura, Y., Hato, E., 2004. Tracking Survey for Individual Travel Behaviour using Mobile Communication Instruments. *Transportation Research Part C: Emerging Technologies*, 12(3-4), 273–291.
- Aurenhammer, F., 1991. Voronoi Diagrams - A Survey of a Fundamental Geometric Data Structure. *ACM Computing Surveys*, 23(3), 345–405.
- Avis, D., Bremner, D., Seidel, R., 1997. How Good are Convex Hull Algorithms? *Computational Geometry*, 7(5&6), 265–301.
- Bar-Gera, H., 2007. Evaluation of a Cellular Phone-based System for Measurements of Traffic Speeds and Travel Times: A Case Study from Israel. *Transportation Research Part C: Emerging Technologies*, 15(6), 380–391.
- Blumenstock, J.E., 2012. Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda. *Information Technology for Development*, 18(2), 107–125.
- Brakatsoulas, S., Pfoser, D., Salas, R., Wenk, C., 2005. On Map-Matching Vehicle Tracking Data. In: *Proceedings of the 31st International Conference on Very Large Data Bases. VLDB Endowment*, Trondheim, Norway, 853–864.
- Brinkhoff, T., 2002. A Framework for Generating Network-Based Moving Objects. *Geoinformatica*, 6(2), 153–180.
- Buchin, K., Buchin, M., Gudmundsson, J., 2010. Constrained Free Space Diagrams: A Tool for Trajectory Analysis. *International Journal of Geographical Information Science*, 24(7), 1101–1125.
- Buchin, K., Buchin, M., van Kreveld, M., Luo, J., 2011. Finding Long and Similar Parts of Trajectories. *Computational Geometry*, 44(9), 465–476.
- Buchin, K., Buchin, M., Wang, Y., 2009. Exact Algorithms for Partial Curve Matching via the Fréchet Distance. In: *SODA '09 Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. 645–654.
- Caceres, N., Wideberg, J.P., Benitez, F.G., 2007. Deriving Origin – Destination Data from a Mobile Phone Network. *IET Intell. Transport. Syst.*, 1(1), 15–26.
- Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C., 2011. Real-Time Urban Monitoring using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems*, 12(1), 141–151.

- Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.-L., 2008. Uncovering Individual and Collective Human Dynamics from Mobile Phone Records. *Journal of Physics A: Mathematical and Theoretical*, 41(22), 1–11.
- Cao, L., Krumm, J., 2009. From GPS Traces to a Routable Road Map. In: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM Press, New York, USA, 3–12.
- Chawathe, S.S., 2007. Segment-Based Map Matching. *IEEE Intelligent Vehicles Symposium*, 1190–1197.
- Crucitti, P., Latora, V., Porta, S., 2006. Centrality in Networks of Urban Streets. *Chaos, Quarterly of the American Institute of Physics*, 16, 1–9.
- Crucitti, P., Latora, V., Porta, S., 2008. Centrality Measures in Spatial Networks of Urban Streets. *Physical Review E*, 73(3), 1–4.
- Csáji, B.C., Browet, A., Traag, V. a., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., Blondel, V.D., 2013. Exploring the Mobility of Mobile Phone Users. *Physica A: Statistical Mechanics and its Applications*, 392(6), 1459–1473.
- De Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., Blondel, V.D., 2013. Unique in the Crowd: The Privacy Bounds of Human Mobility. *Scientific reports*, 3, 1–5.
- Dijkstra, E.W., 1959. A Note on Two Probles in Connexion with Graphs. *Numerische Mathematik*, 1, 269–271.
- Dodge, S., 2011. *Exploring Movement Using Similarity Analysis*. Information visualization. University of Zurich.
- Dodge, S., Laube, P., Weibel, R., 2012. Movement Similarity Assessment Using Symbolic Representation of Trajectories. *International Journal of Geographical Information Science*, 26(9), 1563–1588.
- Dorogovtsev, S.N., Mendes, J.F.F., 2001. Evolution of Networks. *Advances in Physics*, 51(4), 1079–1187.
- Doyle, J., Hung, P., Kelly, D., Farrell, R., 2011. Utilising Mobile Phone Billing Records for Travel Mode Discovery. In: *ISSC*. Trinity College Dublin, Ireland.
- Duckham, M., Kulik, L., 2003. “Simplest” Paths: Automated Route Selection for Navigation. In: *Spatial Information Theory. Foundations of Geographic Information Science*. Springer Berlin Heidelberg, 169–185.
- Eagle, N., de Montjoye, Y.-A., Bettencourt, L.M. a., 2009a. Community Computing: Comparisons between Rural and Urban Societies using Mobile Phone Data. In: *International Conference on Computational Science and Engineering. CSE’09. Ieee*, 144–150.
- Eagle, N., Pentland, A.S., Lazer, D., 2009b. Inferring Friendship Network Structure by using Mobile Phone Data. *Proceedings of the National Academy of Sciences of the United States of America*, 106(36), 15274–8.
- Eclipse, 2014. Eclipse. Retrieved from: <https://www.eclipse.org/> (accessed 28.4.2014).

- Efron, B., 1965. The Convex Hull of a Random Set of Points. *Biometrika*, 52(3&4), 331–343.
- ESRI, 2014. Environmental System Research Institute. Retrieved from: www.esri.com (accessed 8.5.2014).
- European Parliament, 2002. DIRECTIVE 2002/58/EC (Directive on Privacy and Electronic Communications). Retrieved from: <http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32002L0058> (accessed 14.5.2014).
- Faloutsos, C., Jagadish, H.V., Mendelzon, A.O., Milo, T., 1997. A Signature Technique for Similarity-based Queries. In: *Proceedings on Compression and Complexity of Sequences*. IEEE Comput. Soc, 2–20.
- Frias-Martinez, V., Virseda, J., Frias-Martinez, E., 2010. Socio-Economic Levels and Human Mobility. In: *Qual Meets Quand Workshop-QMQ*. Madrid, Spain, 1–6.
- Fu, L., Sun, D., Rilett, L.R., 2006. Heuristic Shortest Path Algorithms for Transportation Applications: State of the Art. *Computers & Operations Research*, 33(11), 3324–3343.
- Furletti, B., Gabrielli, L., Rinzivillo, S., 2012. Identifying Users Profiles from Mobile Calls Habits. *UrbComp'12*, 17–24.
- Geisberger, R., Sanders, P., Schultes, D., Delling, D., 2008. Contraction Hierarchies: Faster and Simpler Hierarchical Routing in Road Networks. In: *Experimental Algorithms*. Springer Berlin Heidelberg, 319–333.
- Geofabrik, 2014. Geofabrik. Retrieved from: www.geofabrik.de/en/ (accessed 10.4.2014).
- GeoTools, 2014. GeoTools. Retrieved from: <http://docs.codehaus.org/display/GEOTOOLS/Home> (accessed 17.1.2014).
- Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., 2007. Trajectory Pattern Mining, 330–339.
- González, M.C., Barabási, A., 2007. From Data to Models. *Nature Physics*, 3, 224–225.
- González, M.C., Hidalgo, C. a, Barabási, A.-L., 2008. Understanding Individual Human Mobility Patterns. *Nature*, 453(7196), 779–782.
- Graser, A., Straub, M., 2013. Ein systematischer Vergleich der Straßennetzwerke von GIP und OpenStreetMap im Großraum Wien. In: *Strobl, J., Blaschke, T., Griesebner, G., Zagel, B. (Eds.), Angewandte Geoinformatik*. Herbert Wichmann Verlag, Berlin/Offenbach, 424–433.
- Gudmundsson, J., Laube, P., Wolle, T., 2008. Movement Patterns in Spatio-Temporal Data. Shekhar S. and Xiong H., eds. *Encyclopedia of GIS*. Berlin: Springer, 726–732.
- Guerra, C., Pascucci, V., 2005. Line-based Object Recognition using Hausdorff distance: from Range Images to Molecular Secondary Structures. *Image and Vision Computing*, 23(4), 405–415.
- Haklay, M., 2010. How good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design*, 37(4), 682–703.

- Haklay, M. (Muki), Weber, P., 2008. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, 7(4), 12–18.
- Hart, P.E., Nilsson, N.J., Raphael, B., 1968. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Transactions on system science and cybernetics*, (4), 100–107.
- Herring, R., Hofleitner, A., Abbeel, P., Bayen, A., 2010. Estimating Arterial Traffic Conditions using Sparse Probe Data. 13th International IEEE Conference on Intelligent Transportation Systems, 929–936.
- Hidalgo, C. a., Rodriguez-Sickert, C., 2008. The Dynamics of a Mobile Phone Network. *Physica A: Statistical Mechanics and its Applications*, 387(12), 3017–3024.
- Järv, O., Ahas, R., Saluveer, E., Derudder, B., Witlox, F., 2012. Mobile Phones in a Traffic Flow: A Geographical Perspective to Evening Rush Hour Traffic Analysis Using Call Detail Records. *PLoS ONE*, 7(11), e49171.
- Järv, O., Ahas, R., Witlox, F., 2014. Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, 38, 122–135.
- Java, 2014. Java. Retrieved from: www.java.com (accessed 8.5.2014).
- Jiang, B., Claramunt, C., 2004. A Structural Approach to the Model Generalization of an Urban Street Network. *Geoinformatica*, 8(2), 157–171.
- Jiang, B., Harrie, L., 2004. Selection of Streets from a Network Using Self-Organizing Maps. *Transactions in GIS*, 8(3), 335–350.
- Jiang, B., Jia, T., 2009. Agent-based Simulation of Human Movement Shaped by the Underlying Street Structure. *International Journal of Geographical Information Science*, 25(1), 51–64.
- Kracht, M., 2004. Tracking and Interviewing Individuals with GPS and GSM Technology on Mobile Electronic Devices. In: 7th International Conference on Travel Survey Methods. Costa Rica, 1–14.
- Krumm, J., Letchner, J., Horvitz, E., 2007. Map Matching with Travel Time Constraints. In: SAE World Congress.
- Kuijpers, B., Miller, H.J., Neutens, T., Othman, W., 2010. Anchor Uncertainty and Space-Time Prisms on Road Networks. *International Journal of Geographical Information Science*, 24(8), 1223–1248.
- Kuter, N., Kuter, S., 2010. Accuracy Comparison between GPS and DGPS: A Field Study at METU Campus. *Italian Journal of Remote Sensing*, 42(3), 3–14.
- Latora, V., Marchiori, M., 2007. A Measure of Centrality based on Network Efficiency. *New Journal of Physics*, 9(6), 188–188.
- Laube, P., Duckham, M., Palaniswami, M., 2011. Deferred Decentralized Movement Pattern Mining for Geosensor Networks. *International Journal of Geographical Information Science*, 25(2), 273–292.

- Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., Huang, Y., 2009. Map-Matching for Low-Sampling-Rate GPS Trajectories. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09. ACM Press, New York, USA, 352–361.
- Maven, 2014. Maven. Retrieved from: <http://maven.apache.org/> (accessed 8.5.2014).
- MDGS, 2014. Millenium Development Goals Indicators - United Nations. Retrieved from: <http://mdgs.un.org/unsd/mdg/> (accessed 29.5.2014).
- Michael, K., McNamee, A., Michael, M.G., Tootell, H., 2006. Location-Based Intelligence – Modeling Behavior in Humans using GPS. In: Technology and Society, 2006. ISTAS 2006. IEEE International Symposium on Technology and Society. 8–11.
- Microsoft, 2014. Microsoft Office Excel. Retrieved from: <http://office.microsoft.com/excel/> (accessed 8.5.2014).
- MobilityLab, 2014. Mobility Lab. Retrieved from: <http://mobilitylab.ut.ee/> (accessed 10.4.2014).
- Mondzech, J., Sester, M., 2011. Quality Analysis of OpenStreetMap Data Based on Application Needs. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 46(2), 115–125.
- Montoliu, R., Gatica-Perez, D., 2010. Discovering Human Places of Interest from Multimodal Mobile Phone Data. Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia - MUM '10, 1–10.
- Mountain, D., Raper, J., 2001a. Modelling Human Spatio-Temporal Behaviour: A Challenge for Location-Based Services. In: Proceedings AGI. Brisbane.
- Mountain, D., Raper, J., 2001b. Positioning Techniques for Location-Based Services (LBS): Characteristics and Limitations of Proposed Solutions. *Aslib Proceedings*, 53(10), 404–412.
- Nanni, M., Pedreschi, D., 2006. Time-focused Clustering of Trajectories of Moving Objects. *Journal of Intelligent Information Systems*, 27(3), 267–289.
- Newson, P., Krumm, J., 2009. Hidden Markov Map Matching through Noise and Sparseness. Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09, 336.
- Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., Barabási, A.-L., 2007. Structure and Tie Strengths in Mobile Communication Networks. *Proceedings of the National Academy of Sciences of the United States of America*, 104(18), 7332–6.
- OSM, 2014. OpenStreetMap. Retrieved from: <http://www.openstreetmap.org> (accessed 10.4.2014).
- Palla, G., Barabási, A.-L., Vicsek, T., 2007. Quantifying Social Group Evolution. *Nature*, 446(7136), 664–7.

- Pelekis, N., Andrienko, G., Andrienko, N., Kopanakis, I., Marketos, G., Theodoridis, Y., 2011. Visually Exploring Movement Data via Similarity-based Analysis. *Journal of Intelligent Information Systems*, 38(2), 343–391.
- Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., Ratti, C., 2010. Activity-aware Map: Identifying Human Daily Activity Pattern using Mobile Phone Data. In: Salah, A.A., Gevers, T., Sebe, N., Vinciarelli, A. (Eds.), *Human Behavior Understanding, Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 14–25.
- Porta, S., Crucitti, P., Latora, V., 2006. The Network Analysis of Urban Streets: A Dual Approach. *Physica A: Statistical Mechanics and its Applications*, 369(2), 853–866.
- Positium, 2014. Positium LBS. Retrieved from: <http://www.positium.ee/> (accessed 10.4.2014).
- Promnoi, S., Tangamchit, P., Pattara-Atikom, W., 2009. Road Traffic Estimation with Signal Matching in Mobile Phone using Large-Size Database. 2009 12th International IEEE Conference on Intelligent Transportation Systems, 1–6.
- Quddus, M. a., Ochieng, W.Y., Noland, R.B., 2007. Current Map-Matching Algorithms for Transport Applications: State-of-the Art and Future Research Directions. *Transportation Research Part C: Emerging Technologies*, 15(5), 312–328.
- Quddus, M.A., Ochieng, W.Y., Zhao, L., Noland, R.B., 2003. A General Map Matching Algorithm for Transport Telematics Applications. *GPS Solutions*, 7(3), 157–167.
- Rahmani, M., Koutsopoulos, H.N., 2013. Path Inference from Sparse Floating Car Data for Urban Networks. *Transportation Research Part C: Emerging Technologies*, 30, 41–54.
- Ratti, C., Pulselli, R.M., Williams, S., Frenchman, D., 2006. Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis. *Environment and Planning B: Planning and Design*, 33(5), 727–748.
- Raubal, M., Miller, H.J., Bridwell, S., 2004. User-Centred Time Geography for Location-Based Services. *Geografiska Annaler, Series B: Human Geography*, 86(4), 245–265.
- Reka, A., Barabási, A.-L., 2002. Statistical Mechanics of Complex Networks. *Reviews of Modern Physics*, 74(1), 47–96.
- Richter, K.-F., Schmid, F., Laube, P., 2012. Semantic Trajectory Compression: Representing Urban Movement in a Nutshell. *Journal of Spatial Information Science*, 4(4), 3–30.
- Rose, G., 2006. Mobile Phones as Traffic Probes: Practices, Prospects and Issues. *Transport Reviews*, 26(3), 275–291.
- RouteWare, 2014. RouteWare. Retrieved from: <http://www.routeware.dk/download.php> (accessed 10.4.2014).
- R-project, 2014. R-project. Retrieved from: <http://www.r-project.org/> (accessed 8.5.2014).
- RStudio, 2014. RStudio. Retrieved from: <https://www.rstudio.com/> (accessed 8.5.2014).

- Rykiel, E.J., 1996. Testing Ecological Models: The Meaning of Validation. *Ecological Modelling*, 90, 229–244.
- Saluveer, E., Ahas, R., 2014. Using Call Detail Records of Mobile Network Operators for Transportation Studies. In: Rasouli, S., Timmermans, H. (Eds.), *Mobile Technologies for Active-Travel Data Collection and Analysis*.
- Shoval, N., Isaacson, M., 2007. Tracking Tourists in the Digital Age. *Annals of Tourism Research*, 34(1), 141–159.
- Sigrist, P., Coppin, P., Hermy, M., 1999. Impact of Forest Canopy on Quality and Accuracy of GPS Measurements. *International Journal of Remote Sensing*, 20(18), 3595–3610.
- Smoreda, Z., Olteanu-Raimond, A.-M., Couronné, T., 2013. Spatiotemporal Data from Mobile Phones for Personal Mobility Assessment. In: *Transport Survey Methods: Best Practice for Decision Making*. Emerald Group Publishing, London, 1–20.
- Song, C., Qu, Z., Blumm, N., Barabási, A.-L., 2010. Limits of Predictability in Human Mobility. *Science*, 327, 1018–1021.
- Thales, S.A., 2005. *Magellan eXplorist 500 - Reference Manual*.
- TNS, 2014. TNS Emor. Retrieved from: <http://www.emor.ee/> (accessed 30.5.2014).
- Toomet, O., Siiri, S., Saluveer, E., Tammaru, T., 2011. Ethnic Segregation in Residence, Work, and Free-time - Evidence from Mobile Communication. Tallinn, Estonia.
- Velaga, N.R., Quddus, M. a., Bristow, A.L., 2009. Developing an Enhanced Weight-based Topological Map-Matching Algorithm for Intelligent Transport Systems. *Transportation Research Part C: Emerging Technologies*, 17(6), 672–683.
- Vlachos, M., Gunopulos, D., Das, G., 2004. Rotation Invariant Distance Measures for Trajectories. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04*, 707.
- Vlachos, M., Kollios, G., Gunopulos, D., 2002. Discovering Similar Multidimensional Trajectories. In: *Proceedings 18th International Conference on Data Engineering*. IEEE Comput. Soc, 673–684.
- Waadt, A., Wang, S., Bruck, G., Jung, P., 2009. Traffic Congestion Estimation Service Exploiting Mobile Assisted Positioning Schemes in GSM Networks. In: *Procedia Earth and Planetary Science 1*. Elsevier B.V., 1385–1392.
- Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabási, A.-L., 2011. Human Mobility, Social Ties, and Link Prediction. In: *KDD'11*. San Diego, California, USA, 1100–1108.
- Wang, H., Calabrese, F., Di Lorenzo, G., Ratti, C., 2010. Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records. *13th International IEEE Conference on Intelligent Transportation Systems*, 318–323.
- Wei, S., 2008. Building Boundary Extraction Base on Lidar Point Clouds Data. In: *Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 37*. 157–161.

- Wentz, E.A., Campbell, A.F., Houston, R., 2003. A Comparison of two Methods to Create Tracks of Moving Objects: Linear Weighted Distance and Constrained Random Walk. *International Journal of Geographical Information Science*, 17(7), 623–645.
- White, C.E., Bernstein, D., Kornhauser, A.L., 2000. Some Map Matching Algorithms for Personal Navigation Assistants. *Transportation Research Part C*, 8, 91–108.
- Wing, M.G., Eklund, A., 2007. Performance Comparison of a Low-Cost Mapping Grade Global Positioning Systems (GPS) Receiver and Consumer Grade GPS Receiver under Dense Forest Canopy. *Journal of Forestry*, 105(1), 9–14.
- Winter, S., Kealy, A., 2012. An Alternative View of Positioning Observations from Low Cost Sensors. *Computers, Environment and Urban Systems*, 36(2), 109–117.
- Work, D.B., Tossavainen, O.-P., Jacobson, Q., Bayen, A.M., 2009. Lagrangian Sensing: Traffic Estimation with Mobile Devices. In: *American Control Conference*. Hyatt Regency Riverfront, St. Louis, MO, USA, 1536–1543.
- Yin, H., Wolfson, O., 2004. A Weight-based Map Matching Method in Moving Objects Databases. *Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM'04)*, 437–438.
- Yuan, Y., Raubal, M., Liu, Y., 2012. Correlating Mobile Phone Usage and Travel Behavior – A Case Study of Harbin, China. *Computers, Environment and Urban Systems*, 36(2), 118–130.
- Zang, H., Baccelli, F., Bolot, J., 2010. Bayesian Inference for Localization in Cellular Networks. *2010 Proceedings IEEE INFOCOM*, 1–9.
- Zhan, F.B., 1997. Three Shortest Path Algorithms on Real Road Networks: Data Structures and Procedures. *Journal of Geographic Information and Decision Analysis*, 1(1), 70–82.
- Zuo, X., Zhang, Y., Feng, C., 2012. A Compute Method of Road Travel Speed based on Mobile Phone Handover Location. *Journal of Networks*, 7(10), 1639–1645.

Personal declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

June 30, 2014

Michelle Fillekes