University of Zurich[UZH]

Department of Geography

Tourist or Commuter? User Characterization based on Spatio-Temporal Footprints in the Absence of Ground Truth

GEO 620 Master's Thesis

Author: Luca Scherrer (11-707-569)
Date of Submission: April 21, 2017

Supervisors:
Dr. Martin Tomko, University of Melbourne
Dr. Peter Ranacher, University of Zurich
Prof. Dr. Robert Weibel, University of Zurich

Faculty Representative:
Prof. Dr. Robert Weibel

Department of Geography, University of Zurich

Contact

Author

Luca Scherrer
Blumenaustrasse 20
8400 Winterthur – Switzerland
Lscherrer7@gmail.com

Supervisors

Dr. Martin Tomko
Geomatics Group
Department of Infrastructure Engineering
The University of Melbourne
Parkville VIC 3010
Melbourne – Australia
tomkom@unimelb.edu.au

Dr. Peter Ranacher
Geographic Information Science (GIS)
Department of Geography
University of Zurich
Winterthurerstrasse 190
8057 Zurich – Switzerland
peter.ranacher@geo.uzh.ch

Prof. Dr. Robert Weibel
Geographic Information Science (GIS)
Department of Geography
University of Zurich
Winterthurerstrasse 190
8057 Zurich – Switzerland
robert.weibel@geo.uzh.ch

Acknowledgements

I am now approaching the end of my studies at the Geography Department of the University of Zurich as well as the end of my master thesis. I would therefore like to express my gratitude to all the people supporting me during my studies and the work on the thesis itself.

I would like to address special thanks to:

- Dr. Martin Tomko, my main supervisor, for all the support, critical comments, and encouragement via Skype from the other side of the world
- Dr. Peter Ranacher, my co-supervisor, for the very generous support, creative ideas, great discussions, and constant guidance on the research field of human movement analysis
- Prof. Dr. Robert Weibel, my co-supervisor and faculty member, for providing important feedback and fruitful discussions
- Sygic, for providing me with great human mobility data
- Jörg Roth from GIUZ IT, for all the help setting up my Server
- Dr. Curdin Derungs, Dr. Michele Volpi, Oliver Burkhard and Michelle Fillekes, for the helpful hints and fruitful discussions in many ways
- My dear friends, especially Joel Durand and Kevin Meier, for proofreading my thesis and being there when needed
- A huge thanks to my family and especially to Anja, for their continuous support during my studies

Thank you!

Luca Scherrer
University of Zurich
April 2017

Abstract

Our individual spatio-temporal behaviors can be captured by various sensors embedded in the personal devices we carry, and by the environments we visit. These recordings can originate from various sources, such as GPS sensors on mobile phones, phone calls, or the purchase and usage of metro tickets (Hasan et al. 2013). Movement data that originates from such sensors often have a high resolution in both space and time, however, are often lacking additional knowledge of the surveyed community. In contrast to more traditional data sources such as census or interview data, such novel data types offer an actual tracking of the people's behavior in space and time.

The growing amount of human movement data provides us with both new opportunities and newly emerging challenges in human mobility research. Accordingly, considerable research has been conducted in order to find patterns in human mobility. In various studies of recent years (González et al. 2008; Palchykov et al. 2014; Calabrese et al. 2013), the surveyed people, however, have often been considered members of one large, homogeneous community. This is a prevalent approach, since GPS or call detail records (CDR) often lack ground truth, i.e. no information about the true membership of the individual users to a certain community are available.

The aim of this thesis is therefore to overcome the unavailability of ground truth, by developing a methodology to categorize users into different user types based on their spatio-temporal footprints. The methodology consists of a series unsupervised machine learning techniques (principal component analysis, clustering) and will be applied using one month of data from a navigation app over the whole of Australia. We further present a set of methods to analyze the preferred visit locations and the temporal patterns of these visits for the most dominant user types found in the two biggest Australian cities, Sydney, and Melbourne. Based on these methods, we show that distinct elaborated user types such as tourists or commuters visit areas with different likelihoods and magnitudes. Accordingly, we are presented with a deeper understanding of the spatio-temporal dynamics of different user types and their preferred visit locations, that cannot be found in traditional surveys.

Zusammenfassung

Unser individuelles raumzeitliches Verhalten kann von verschiedensten Sensoren, eingebettet in unsere persönlichen Geräte, sowie von der Umgebung welche wir besuchen, eingefangen werden. Diesee Aufnahmen können von verschiedensten Quellen stammen, etwa von GPS Sensoren in unseren Mobiltelefonen, von Telefonanrufen, oder etwa vom Kauf und der Nutzung von Metrotickets (Hasan et al. 2013). Bewegungsdaten, welche von solchen Sensoren stammen, haben oft eine hohe räumliche und zeitliche Auflösung, ihnen mangelt es jedoch an zusätzlichen Informationen bezüglich der untersuchten Gemeinschaft. Solche neuartigen Datenquellen bieten jedoch, im Gegensatz zu traditionelleren Datenquellen wie etwa Zensus oder Interviewdaten, ein tatsächliches Tracking des menschlichen Verhaltens in Raum und Zeit.

Die wachsende Anzahl an Bewegungsdaten verschafft uns sowohl neue Möglichkeiten wie auch neu aufkommende Herausforderungen in der Forschung von menschlicher Mobilität. Eine beträchtliche Anzahl Studien wurde dementsprechend ausgeführt um Muster in menschlicher Mobilität zu entdecken. In mehreren Studien der letzten Jahre (González et al. 2008; Palchykov et al. 2014; Calabrese et al. 2013), wurden jedoch die untersuchten Personen jeweils als eine homogene Gemeinschaft dargestellt. Das ist ein gängiger Vorgang, da GPS-Daten oder Telefonverbindungsdaten (CDR) häufig der sogenannte Ground Truth fehlt, d.h. keine Informationen über die wirklichen Zugehörigkeiten von einzelnen Nutzern sind vorhanden.

Das Ziel dieser Arbeit ist es dementsprechend, das Problem des Nichtvorhandenseins von Ground Truth zu überwinden, indem eine Methodik präsentiert wird, welche Nutzer in unterschiedliche Nutzertypen charakterisiert, basierend auf deren raumzeitlichen Profilen. Die Methodik besteht aus einer Folge von unbewachten Machine Learning Techniken und wird auf Daten eines Navigations-Apps über einen Monat und ganz Australien angewendet. Wir präsentieren zudem eine Zusammenstellung an Methoden, um die besuchten Orte sowie die zeitlichen Muster der dominanten Nutzertypen in den zwei Städten Melbourne und Sydney zu untersuchen. Ausgehend von diesen Methoden zeigen wir, das unterschiedliche ausgearbeitete Nutzertypen wie etwa Touristen oder Pendler, Gebiete mit unterschiedlichen Wahrscheinlichkeiten und Magnituden besuchen. Dementsprechend erhalten wir ein tieferes Verständnis der raumzeitlichen Dynamiken der unterschiedlichen Nutzertypen sowie deren bevorzugten besuchten Orten, welches wir nicht durch traditionelle Studien erhalten.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABS:	Australian Bureau of Statistics
AD:	Average Distance
ADM:	Average Distance between Means
AGNES:	Agglomerative Nesting
AIC:	Akiake Information Criterion
API:	Application Programming Interface
APN:	Average Proportion of Non-overlap
CBD:	Central Business District
CDR:	Call Detail Records
CE:	Cross-Entropy Monte Carlo Algorithm
CLARA:	Clustering Large Applications
Def.	Definition
DIANA:	Divisive Analysis
DTW:	Dynamic Time Warping
ER-Model:	Entity-Relation Model
FFT:	Fast Fourier Transformation
GPS:	Global Positioning System
ID:	Identifier
iOS:	Apple's mobile operating system, formerly known as iPhone OS
IQR:	Interquartile Range
KDD:	Knowledge Discovery in Databases
ODM:	Origin-Destination Matrix
OSM:	OpenStreetMap
PAM:	Partitioning Around Medoids
PCA:	Principal Component Analysis
PC*n*:	Principal Component Number *n*
POI:	Point of Interest
RFID:	Radio-Frequency Identification
SA1:	Statistical Areas Level 1
SA2:	Statistical Areas Level 2
SQL:	Structured Query Language
SSE:	Start, Stop or End Point of a trajectory
Std. Dev.:	Standard Deviation
UTC:	Coordinated Universal Time
WKT:	Weighted Kendall's Tau distance
WSF:	Weighted Spearman Footrule distance
3PC-X3KM:	3 K-Means Clusters based on 3 Principle Components
4PC-X5KM:	5 K-Means Clusters based on 4 Principle Components

1. Introduction

1.1 Context and Motivation

Our individual spatio-temporal behaviors can be captured by various sensors embedded in the personal devices we carry, and by the environments we visit. These recordings can originate from various sources, such as GPS sensors on mobile phones, phone calls, or the purchase and usage of metro tickets (Hasan et al. 2013).

While the numbers of mobile phones and their usage increase steadily, the amount of sensor data itself that captures human movement is increasing as well (Parent et al. 2013; Zheng et al. 2013). Technological innovations further enable us to easily store big amounts of sensor data (Han et al. 2012).

The emergence of data that captures human movement provides us with both new research opportunities and newly emerging challenges. According to Zook et al. (2015), location-based data such as mobility data can give insight into human movement in both space and time. The challenge for researchers is to benefit from the new possibilities that arise from the increase in data and the evolution of new methods and opportunities. Human mobility data are therefore of great interest to various groups of researchers, including tourism researchers (Ahas et al. 2008; Edwards & Griffin 2013; Shoval et al. 2011), transport and urban planners (Ahas et al. 2015; Noulas et al. 2012; Yuan et al. 2012; Yuan & Raubal 2012), demographers, and others.

An interpretation of the movement itself, however, is often not directly interpretable solely based on raw movement trajectories, i.e. the raw timestamped coordinates as recorded by the device and stored in a database. To gain information out of these recorded movements and to achieve a deeper understanding of human mobility itself, it is crucial to analyze these raw data sets with suitable methods, in order to extract interpretable behavioral patterns (Renso et al. 2012).

The processing and analysis of human mobility data, however, are often not feasible with simple data processing techniques due to the size and variability of the data. A possible

solution is to work with databases and data mining techniques. Using data mining in combination with movement data is a field of research on its own, called movement mining and "*aims for the conceptualizing and the detecting of non-random properties and relationships in movement data that are valid, novel, useful, and ultimately understandable*" (Laube 2014, p.31).

This thesis will contribute to this domain by analyzing a large amount of raw movement data that is not directly interpretable by humans. Accordingly, we process it in such a way that it becomes interpretable and we show how the interpreted data can reveal potentially interesting and previously hidden patterns in urban space.

1.2 Problem Statement

Researchers studying mobility have used several types of data such as interviews, observations, or census statistics. Each of these data types about mobility behavior is different. Census data presents a snapshot over a large sample size, but no longitudinal data. In contrast, data from interviews and observations often offer longitudinal data, but only over a small surveyed community. In all cases however, the data is mostly enriched with additional knowledge about the demographic characteristics of the surveyed community.

Novel movement data sources with a higher resolution in both space and time, however, are often lacking additional knowledge of the surveyed community. In contrast to census or interview data, they offer movement tracking of people in space and time, allowing it to record people's actual behavior. Accordingly, there is much more information about movement in novel data sources than in census data or observational data.

One of these novel data source that has been widely used in the last two decades is call detail records (CDR), i.e. data generated by mobile phone communication activities. Due to their nature (see **section 2.1.1**), CDR offer data for a large amount of people in a relatively high spatial and temporal resolution. Ahas et al. (2007) and Yuan & Raubal (2012), for example, used these data to identify regions with distinct visiting patterns. Others such as Candia et al. (2008) and Jiang et al. (2013) used CDR to understand collective human behavior and mobility networks.

As Zhang et al. (2014) and Zhao et al. (2016) state, however, CDR also has its drawbacks, since it is biased and therefore not always an optimal type of data for mobility research. This bias is caused by the fact that not all types of people (residents, tourists, etc.) use their mobile phones in the same way and generate the same amount of data. Moreover, people tend to contact other people at specific places such as home or work. The places extracted from these data then often only cover a small amount of all visited places (Zhao et al. 2016).

A different type of data used for human mobility research in recent years is global positioning system (GPS) data. GPS data offers a high spatial resolution and mostly, also a high temporal resolution. Accordingly, it is generated and collected even in vast amounts. The high spatio-temporal resolution generates privacy conflicts, since small amounts of GPS recordings permits the unique identification of individuals (de Montjoye

et al. 2013). For researchers, GPS data of a large amount of people with both a high spatial and temporal resolution are therefore not easy to obtain. Due to that, we use a type of data in this thesis that combines the advantages of both the CDR and the GPS data.

In human mobility research, surveyed people are often considered members of one large, homogeneous community. Although arguing that they are looking at individual trajectories and users, González et al. (2008) state that all *humans follow simple and reproducible patterns*. Others (Palchykov et al. 2014; Calabrese et al. 2013) speak of individual mobility, but nevertheless treat all users as one homogeneous community. The possibility that the data represent trajectories of different types of users visiting different places and areas is often neglected. This is a prevalent approach, since GPS or CDR often lack ground truth, i.e. no information about the true membership of the individual users to a certain community is available.

In summary, traditional surveys that can be used for mobility research offer a wide range of information about the surveyed community, but lack actual recordings of the movement of people. Novel data sources offer the recordings of people's movement with a high spatial and temporal resolution although, they are biased.

1.3 Research Aims

The aim of this thesis is therefore to overcome the stated problems by using a novel data source that combines the advantages of both CDR and GPS data by having both a high spatial and temporal resolution: movement data captured by a smartphone navigation app (1). Furthermore, we neglect the often-used idea that all users belong to the same homogeneous community. For this reason, we try to categorize users into different user types based on their spatio-temporal behaviors (2). By analyzing the actual areas that the different user types visit, we are presented with a deeper understanding of the spatio-temporal dynamics of different user types in the two biggest Australian cities, Sydney and Melbourne, that cannot be found in traditional surveys (3). We therefore analyze the spatio-temporal behavior and footprints of all users in the whole of Australia, but then test how the distinct groups use the two cities.

To achieve the stated aims of the thesis, we explore two research questions. The first research question aims at the characterization of users into different user types:

Research Question 1:	*How can individual users of the navigation app be characterized based on their spatio-temporal footprints in the absence of ground truth? What are the principal factors describing the different user types?*
Hypothesis 1:	*Navigation app users can be characterized into different user types based on their spatio-temporal footprints and computed mobility patterns.*

For the purpose of this thesis, both the terms user type and spatio-temporal footprint will formally be defined later in **section 3.1**.

The established user types only characterize users based on descriptive characteristics of the geometries of the spatio-temporal footprints, independent of the actual location of the footprints themselves. The fact that a user regularly and daily – except weekends –, moves between a location A and B and back, in the morning and afternoon, relates to a possible work commuting pattern, disregarding whether A and B are in Sydney or Melbourne. Accordingly, we transform groups of similarly acting users into a human-interpretable form, which ultimately leads to the formation of the user types.

In the second research question, we further explore the temporal and spatial characteristics of the identified user types in the two cities of Melbourne and Sydney. This presents us with a deeper understanding of space use by different user types. To address this, the following research question and hypothesis are examined:

Research Question 2:	*What are the spatio-temporal usage patterns of the identified types of users in the two cities of Melbourne and Sydney? Can individual areas be characterized based on temporal usage patterns of different user types?*
Hypothesis 2:	*Different user type use the two investigated cities in different ways. They visit different places and have different temporal usage patterns. Accordingly, different urban areas show distinct visiting patterns by different user types.*

1.4 Main Outcomes

The main contribution of this thesis is a) a methodology to characterize users based on their spatio -temporal patterns into human-interpretable user types in the absence of ground truth, based on unsupervised machine learning. Additionally, b) we present a set of methods to analyze the preferred visit locations and the temporal patterns of these visits for the most dominant user types found.

1.5 Thesis Structure

Following this introduction is **Chapter 2**, which provides an overview of the work related to this thesis. **Chapter 3** gives an overview of the methodology and data used in this thesis as well as the pre-processing carried out. Following that is **Chapter 4**, which presents the data analysis steps carried out to characterize the individual users into different user types. **Chapter 5** then presents an analysis of the spatial and temporal characteristics of the user types found in Melbourne and Sydney. In **Chapter 6**, the carried research questions are discussed in detail and put into context with the existing literature. Fur-

thermore, possible limitations are discussed in this chapter. The thesis finishes with **Chapter 7**, which concludes and summarizes the main aspects and provides an outlook for future work.

2. Background and Related Work

This section serves as a first theoretical introduction into the theme of human mobility research and user characterization. Human mobility research is not a new research field as such, however, the large availability and pervasiveness of location-acquisition techniques as well as the associated newly arisen data formats fueled its increasing emergence in recent years. The emergence of mobile phones in the last 20 years probably had the biggest impact on this trend. The numbers of both the usage as well as the amount of users and devices are increasing steadily, which leads to a higher availability of often huge amounts of location-based data produced by these devices (Parent et al. 2013; Renso et al. 2012; Zheng et al. 2013). Mobile phones are not only ubiquitous in today's life, they also enable researchers to collect data in both high spatial (meters) and temporal resolution (seconds) (Birenboim & Shoval 2015).

Section 2.1 gives an overview of the different movement perspectives as well as the most often used data types used for human mobility research. In section 2.2, an introduction into data mining and movement mining is given. The chapter ends with section 2.6, which provides insight into several studies that deal with different topics involving human mobility research.

2.1 Movement Perspectives and Data Collection Types

A wide range of different data types exist that can be used for mobility research. Thus, mainly two different types of movement perspectives and approaches to collect movement data can be identified; *Lagrangian* and *Eulerian* (Laube 2014; Dodge et al. 2016). While the Lagrangian perspective focuses on the change of position of a moving object, the Eulerian perspective addresses the tracking of objects as they are passing by an observation point (see Figure 2.1).

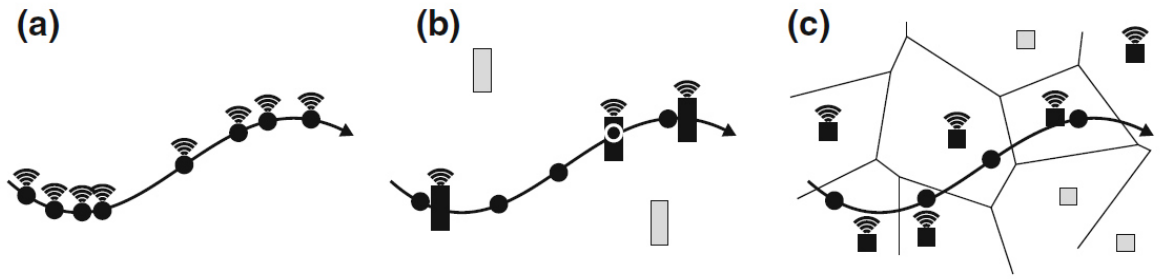

Figure 2.1: Lagrangian vs. Eulerian perspective: (a): Lagrangian perspective (e.g. movement of GPS-tracked animal). (b) Eulerian perspective (e.g. movement along fixed traffic census points). (c) Eulerian perspective (e.g. movement along a series of radio cells) (Laube 2014, p.13)

In this thesis, we mostly address data types originating from mobile phones, since we are working with mobile phone data as well. According to Giannotti et al. (2011), such mobile phone data provide an ideal base to understand human behavior. Mobile phone data can, however, be collected in many ways and there is no *single* typical mobile phone data. In the following sections, we therefore discuss the three dominant types of data sets used for tracking mobile users in space and time.

2.1.1 Call Detail Records

Call detail records (CDR) data are generated by mobile phone communication activities, i.e. making/receiving a call, or sending/receiving a text message. Based on the terminology, call detail records belong to the Eulerian approach of data collection. Movement is collected as the change of the object relative to a fixed point in space, in this case radio antennas (Laube 2014).

CDR data consist of several attributes, including the phone number of both the calling and the receiving user, the start time and the duration of the call (Zheng et al. 2014), as it can be seen in Figure 2.2. The meaning of position in the context of CDR data, however, cannot be compared to position as in GPS-position. The stored position of CDR-data consists of the position of the nearest mobile phone antenna and can therefore only serve as a proxy for the exact position of the phone call. Thus, the accuracy of positioning of a mobile phone is dependent on the seize and shape of a radio cell around a radio antenna.

CDR data have been used for various research topics, i.e. building networks between users (Zheng et al. 2014), examining the variability in human activity spaces (Järv et al. 2014), detecting differences in everyday activities (Ahas et al. 2015) and understanding the spatiotemporal distribution of people within a city (Toole et al. 2012).

user_id	datetime	antenna_id	direction	interaction	Country	Phone type
dsdasdfasdfqwerwoiouiruf9w90283u4oijofjlksdkfansdkgklsjoqi4029830498203	2015.08.30 08:17:47	90	out	text	21465	986745342
f2436657210815553388ed1adacec6e38c1fwewer20ed3d560a4a6bb8e55571	2015.08.30 08:16:06	461	in	call	21464	986745342
efdkssdfeedwerwoiouiruf9w90283u4oijofjlksdkfansdkgklsjoqi4029830sdfasddf	2015.08.30 08:04:32	223	in	call	24134	986745342
7a27adb89209cb27eec47bc79aaasdfwewef612f77760e608fd2ea53c3a74f0c95	2015.08.30 08:40:02	454	out	call	23413	986745342

Figure 2.2: A snapshot of call detail records (Leng et al. 2016, p.2)

2.1.2 Handover data

Handover data is data that is generated when a mobile phone user moves from one radio cell to another. Thus, a mobile phone disconnects from one cell and connects to the other

(Sagl et al. 2012). Accordingly, handover data can be categorized as a Eulerian data collection approach. Examples of handover data used for mobility research includes works by Demissie et al. (2013) who used handover data to detect the traffic status on roads, whereas Sagl et al. (2012) presented a visual analytics approach of handover data to extract collective spatio-temporal mobility.

In research, handover data have not been used as often as other data types for tracking outdoors, mainly due to two reasons. First, not all radio cells are of equal size and shape due to their uneven spatial distribution, which makes mobile positioning difficult. The second reason is related to the so-called *ping-pong-effect* (Vajakas et al. 2015). This effect happens when a mobile phone alternatively connects to various radio antennas nearby. The then arising pattern then might be rather confusing, since it is unknown whether the mobile phone is moving or not.

Handover data are, however, typical for tracking of human movement in indoor spaces, such as shopping malls, where Wi-Fi access logs capture handovers between different access points (Ren et al. 2016).

2.1.3 GPS Tracking Data

The global positioning system (GPS) is made up of a series of satellites orbiting the earth. These satellites emit signals that are picked up by the receivers (e.g. GPS sensors in mobile phones). With at least four satellites in range, it is possible to triangulate the receivers position (Shoval & Isaacson 2007). Due to that and in contrast to the previous discussed positioning techniques, GPS offers a higher spatial accuracy. Accordingly, GPS tracking is a Lagrangian approach of data collection, unlike the previously presented data types.

Another difference to both CDR and handover data lies in the source of the collected GPS data. Whereas CDR and handover are restricted to mobile phones, GPS data can additionally be collected from, for example, on-board car GPS receivers (Giannotti et al. 2011; Pappalardo et al. 2013; Andrienko et al. 2015).

Like CDR or handover data, GPS data has its drawbacks. Firstly, it is very sensitive data type due to the fine position granularity. Secondly, researchers are often not presented with demographic characteristics of the specific users which makes it hard to generate some sort of ground truth.

2.2 Data Mining

Data mining is a research field in computer sciences and deals with the automated extraction of patterns and knowledge from large databases or other large repositories (Han et al. 2012). Data mining, sometimes referred to as *knowledge discovery in databases* (KDD), has gained a lot of attention in recent years, mainly due to the availability of large amounts of data. Han et al. (2012, p.2) remark that the growing amount of data in recent years has additionally lead to the understanding, that "*powerful tools are needed to automatically uncover valuable information and to transform data into organized knowledge*".

Data mining is a process of knowledge acquisition (Han et al. 2012, pp.6–8), consisting of various steps that can be summarized as follows:

- *Data cleaning: the removal of noise and irrelevant data*
- *Data integration: the combination of multiple data sources*
- *Data selection: the retrieval of data relevant to the analysis talk from the database*
- *Data transformation: the transformation of data to have them in an appropriate form for mining*
- *Data mining: the application of intelligent methods to extract data patterns*
- *Pattern evaluation: the identification of truly interesting patterns that are representing knowledge based on interestingness measures*
- *Knowledge presentation: the application of visualization and knowledge representation techniques to present the mined knowledge to the user*

A similar approach is presented by Fayyad et al. (1996) and can be seen in Figure 2.3. Here, data mining is again only one step in a series of processes, ultimately leading to the discovering of useful information and knowledge in the data. The steps here included are: integration of the data from multiple and heterogeneous data sources, selection of useful data, preprocessing, transformation, data mining itself and interpretation and evaluation.

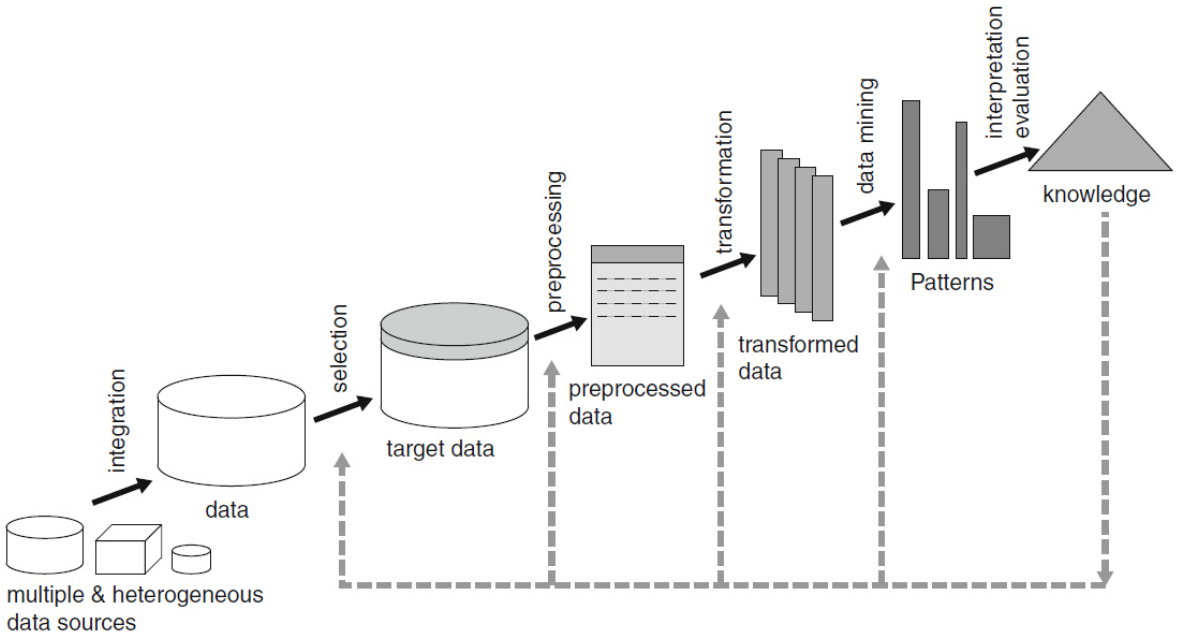

Figure 2.3: Six steps of the KDD process (Laube 2014, p.30; adapted from Fayyad et al. 1996)

The individual steps of the data mining process can therefore be seen as a cyclic process of going back and forth again. This means that a data mining step may not only lead to an interpretation, but also to a re-transformation of the data itself.

2.2.1 Movement Mining

Data mining can be used in combination with movement data. Laube (2014) defines the term used for that combination, *movement mining*, as follows:

> ***Movement Mining:*** *Movement mining aims for conceptualizing and detecting non-random properties and relationships in movement data that are valid, novel, useful, and ultimately understandable* (Laube 2014, p.31).

Thus, Laube (2014, pp.31–32) defines the four terms valid, novel, useful and ultimately understandable as follows:

- valid: *"properties and relationships should be applicable to new data as well"*
- novel: *"properties and relationships should be nontrivial and unexpected"*
- ultimately understandable: *"properties and relationships should be simple and interpretable for domain experts"*
- useful: *"properties and relationships should be useful for further decision making process"*

Laube (2014) further outlines several processes of movement mining, that will also be applied in this thesis: Segmentation and filtering (**section 3.6**), similarity and clustering (**sections 4.2** and **4.3**), movement pattern extraction (i.e. spatio-temporal footprints, **section 4.1**) and exploratory analysis and visualization (**sections 5.1** and **5.2**).

Renso et al. (2012) as well as Zheng (2015) use a similar term that refers to the knowledge discovery in movement databases; *trajectory data mining*. The goal of trajectory data mining is to use data mining techniques to extract mobility patterns from a large number of trajectories (Renso et al. 2012).

2.2.2 Supervised vs. Unsupervised (Machine) Learning

According to Witten et al. (2011), machine learning refers to the technical basis of the a data mining approach. Zhou (2003) sees machine learning and data mining as two separate disciplines, whereas data mining has received a lot of contributions from the machine learning domain. This is further highlighted in a statement about the objective of Witten et al.'s (2011, p.xxiv) book, namely to *"introduce the tools and techniques for machine learning that are used in data mining"*. Accordingly, we can see machine learning both as a separate domain as well as a set of techniques that can be used in a data mining approach.

Most of the learning problems within a data mining or a machine learning approach can be categorized either into a *supervised* or an *unsupervised* problem (Han et al. 2012; James et al. 2013). In supervised learning, unlabeled objects are assigned a class label using a model that has been developed based on objects with a known class label (Tan et al. 2006). Opposite to that is unsupervised learning, in which we apply an algorithm on objects whose class labels are unknown in order to discover classes within the data. The

found and learned model, however, cannot tell us the actual semantic meaning of the classes found, since the training data is not labelled (Han et al. 2012).

In the following three sections, we will give examples of some of the most prominent learning approaches that are also relevant for this thesis.

2.3 Classification

To find groups in a set of data objects such as trajectories or users, several approaches can be made, dependent on the information found in the data. One approach is to categorize objects without a label with the help of some known objects with a class label. The term used for that supervised learning approach is called classification and is defined by Han et al. (2012) as follows:

> ***Classification:*** *The process of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown* (adapted from Han et al. 2012, p.18).

For a classification, we first need a collection of records (training set) with a set of attributes. One of these attributes most consist of the class label itself. The goal of the classification is now to find a model for the class labels, based on the values of the other attributes (Tan et al. 2006). The model can then be applied on a *test set*, whose class labels can be unknown (see Figure 2.4). Essential in the classification approach is that the derived model is based just on a comparatively small set of the total data, called *training set,* and not on the total data. The test set used to test the model is therefore even smaller than the training set. Various classification techniques exist, including decision trees, rule-based methods, memory based reasoning, neural networks or naïve Bayes (Tan et al. 2006; Wu et al. 2011; Han et al. 2012). Cross-validation methods can further be used to assure that the impact of training and test data bias is minimized.

The above definition can further be extended for the case of trajectories, as it is done by Lee et al. (2008, p.1081). They define trajectory classification "*as the process of predicting class labels of moving objects based on their trajectories as well as other* (computed) *features*". Examples of trajectory classifications include Lee et al.'s (2008) work on vessel type classification or Trasarti et al.'s (2011) approach on classifying trajectories to find compatible carpooling users.

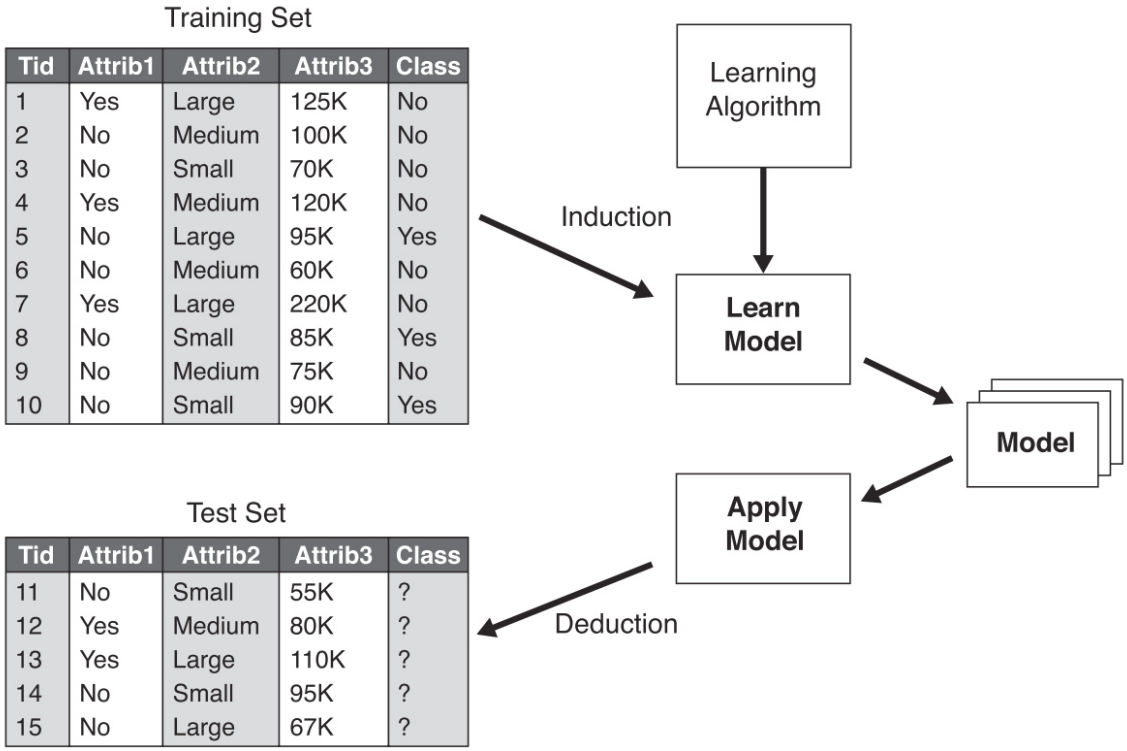

Figure 2.4: An example of a classification approach. First, based on a training set and the help of a learning algorithm, a model is built. The model will then be applied on a test set class labels are unknown (Tan et al. 2006, p.148).

2.4 Clustering

Clustering is an unsupervised machine learning method that can be used to discover groups of similar objects without a prior knowledge of any class labels (Tan et al. 2006). The class labels can, however, be generated after the clustering approach by interpreting the different clusters qualitatively. The intra-cluster distances of clusters found should therefore be minimized whereas the inter-cluster distances should be maximized (Han et al. 2012). Accordingly, the greater the similarity within the group and the greater the differences between the groups, the better the clustering (Tan et al. 2006).

A variety of clustering algorithms exist. An important distinction must be made between partitional and hierarchical set of clusters. The partitional clustering approach is based on the division of the data objects into non-overlapping clusters (Tan et al. 2006). Examples of three different partitional clustering approaches can be seen in Figure 2.5. Hierarchical clustering on the other hand leads to a hierarchical decomposition of the set of data objects (Han et al. 2012). Accordingly, clusters can be identified at multiple scales (Thomason et al. 2016). In exclusive clustering, each object is assigned to a single cluster whereas in overlapping (non-exclusive) clustering, an object can belong to more than one single cluster. A mixture between exclusive and non-overlapping is fuzzy clustering. Here, each object is assigned to each cluster with a certain membership degree between 0 and 1 (Tan et al. 2006). The last distinction lies between complete and partial clustering. In complete clustering, each object is assigned to an object whereas in partial clustering not every object has to be assigned to one cluster.

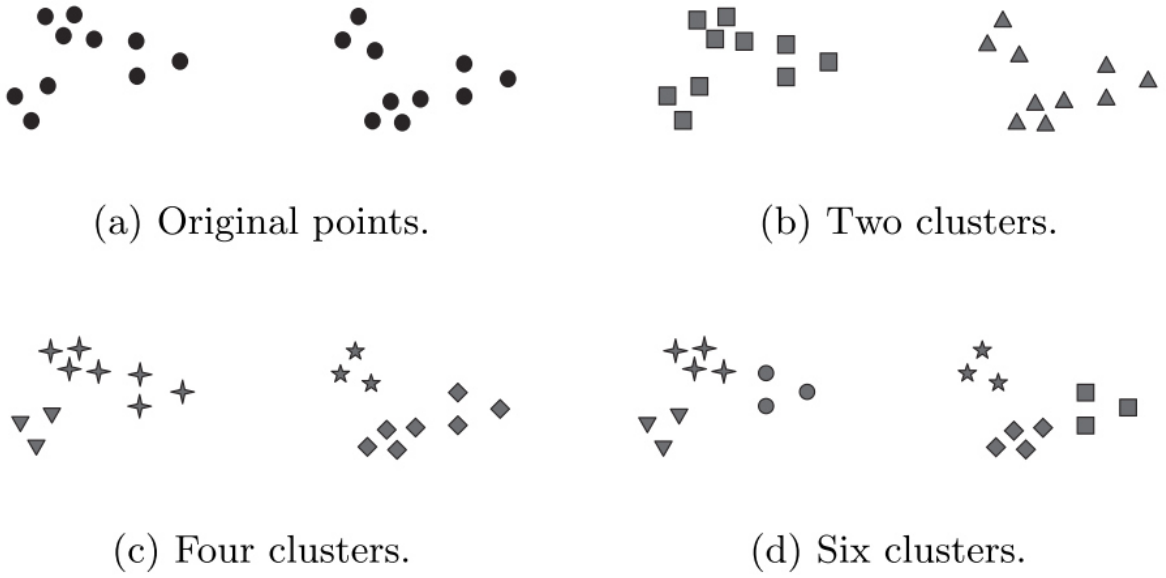

Figure 2.5: Three different ways of clustering based on the same original points (Tan et al. 2006, p.491).

2.4.1 Clustering Algorithms

In this thesis, four different clustering algorithms are used which are shortly presented in the following sub-sections:

K-Means

One of the most known clustering algorithms, K-Means, is a centroid-based clustering technique, which means that it uses the centroid of the clusters to represent the cluster. The quality of K-Means is therefore measured as the within-cluster validation, which is the sum of the squared error between the individual objects in a cluster and their centroid (Jain 2010; Han et al. 2012). To compute the standard error, K-Means randomly choses k objects from the data set as starting centroids, i.e. cluster centers. It then assigns each data object to the cluster it is most similar. After that, it re-chooses the cluster center until the mean value of the cluster is not changing anymore, therefore being smallest as possible. A disadvantage of K-Means is that outliers have a strong influence on the mean value of the clusters, which affects the assignment of other objects to the clusters (Han et al. 2012).

K-Medoid

An alternative to K-Means is presented by the K-Medoids method which does not use a mean value (centroid) to represent the cluster, but an actual existing object. Each remaining object is assigned to the cluster of which the representative object is most similar. The goal of the K-Medoids method is to minimize the absolute-error criterion, which is the sum of dissimilarities between the individual objects and the corresponding representative object (Hastie et al. 2009; Han et al. 2012).

Partitioning Around Medoids (PAM) is an implementation of the K-Medoids method. The objective of PAM is to find the representative object (medoid, most centrally located) for each cluster. Since PAM tests all objects in the data set until it finds the best medoids, it is a time consuming algorithm and therefore does not work well for large data sets (Han

et al. 2012). CLARA (Clustering Large Applications) is an implementation of PAM that only uses subsets of the total data set to find the best clustering (Halkidi et al. 2001).

AGNES

Agglomerative nesting (AGNES) is an agglomerative hierarchical clustering method that initially, places each element into a cluster of its own. The elements are then merged stepwise based on a closeness criterion such as the Euclidean distance (Brock et al. 2008). For example, two clusters are merged if elements of these two form the minimum Euclidean distance between any two objects in the given clusters (Han et al. 2012).

DIANA

Divisive analysis (DIANA) is, as the name already states, a divisive hierarchical clustering method. At the beginning, DIANA starts with all elements in one single cluster and then step by step divides the cluster into smaller clusters until each element belongs to only one cluster (Brock et al. 2008). DIANA splits the cluster based on a criterion such as the maximum Euclidean distance between the closest neighboring elements in a cluster (Han et al. 2012).

2.4.2 Cluster Validation

Determining the right and appropriate number of clusters for a given data set is one of the most difficult problems in data clustering (Jain 2010), since a variety of decisions can have big influences on the result (James et al. 2013). Ideally, the result of a clustering has both good statistical properties as well as useful and interpretable solutions (Brock et al. 2008; James et al. 2013).

The appropriate number of clusters is dependent on various decisions. First, the chosen clustering algorithm and secondly, whether the input variables should be scaled to have a standard deviation of one. By scaling the individual input variables, we specify that each variable will be given the same importance in the clustering (James et al. 2013).

To find both the the accurate clustering algorithm for a given problem as well as the number of clusters k, a variety of methods exists. We therefore present five different methods, for which an overview is given in Table 2.1.

Table 2.1: Summary of different clustering validation methods

Cluster validation method	Value demonstrating good clustering result	Value range
Silhouette width	high	-1,+1
Gap statistic	high	0, ∞
Average proportion of non-overlap (APN)	low	0,1
Average distance (AD)	low	0, ∞
Average distance between means (ADM)	low	0, ∞

Silhouette Width

An often-used approach is calculating the silhouette width for each number of *k.* The value of the silhouette width measures the degree of confidence on how well each element *I* is clustered (Brock et al. 2008). Well clustered elements have a silhouette width value near 1 whereas poor clustered elements have values near -1. Accordingly, the silhouette width should be as close to 1 as possible (Rousseeuw 1987; Brock et al. 2008).

Gap Statistic

Another approach is presented by Tibshirani et al. (2001)'s gap statistic. The gap statistic runs the clustering algorithm for various sizes of *k*. It then calculates the dispersion for each *k,* which is the sum of all distances from the points to their cluster mean. By sampling uniformly from the original data set, *B* reference data sets are then formed, from which the dispersion is calculated as well. The gap for each size of *k* is then defined as the log value of the mean dispersion of a reference data set minus the log value of the dispersion of the original data set.

Stability Measures

Besides the silhouette width and the gap statistics, there are additional metrics to measure the stability of a clustering, such as the *stability measures*. These measures compare the clustering results of the overall data with the results of clusterings made when removing each column, i.e. each data element (Datta & Datta 2003). According to Brock et al. (2008), the measures generate good results especially when the data is highly correlated.

The first of these measures is called average proportion of non-overlap (APN). It measures the average proportion of elements that are not clustered in the same cluster when clustered with an element removed (Brock et al. 2008). APN generates values in the interval [0,1], whereas values close to 0 correspond to consistent clustering results.

The average distance (AD) is a measure that describes the average distance between elements of a cluster when clustering all data and when a column is removed. Unlike APN, AD computes values between 0 and ∞, whereas values as close to 0 as possible generate the best clustering results (Brock et al. 2008).

The average distance between means (ADM) describes the average distance between the cluster centers again based on the total data and when an element is removed. Like AD, ADM generates values between 0 and ∞, whereas small values are preferred (Brock et al. 2008).

2.5 Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised method to summarize a large set of variables to a smaller number of representative variables, i.e. principal components (James et al. 2013). Unlike a clustering method, PCA focuses on the analysis of the similarities and therefore tries to explain as much variation as possible (van den Berg et al. 2006). The newly established principal components (see Figure 2.6) are linear combinations of the original variables, derived from the eigenvectors of the covariance matrix.

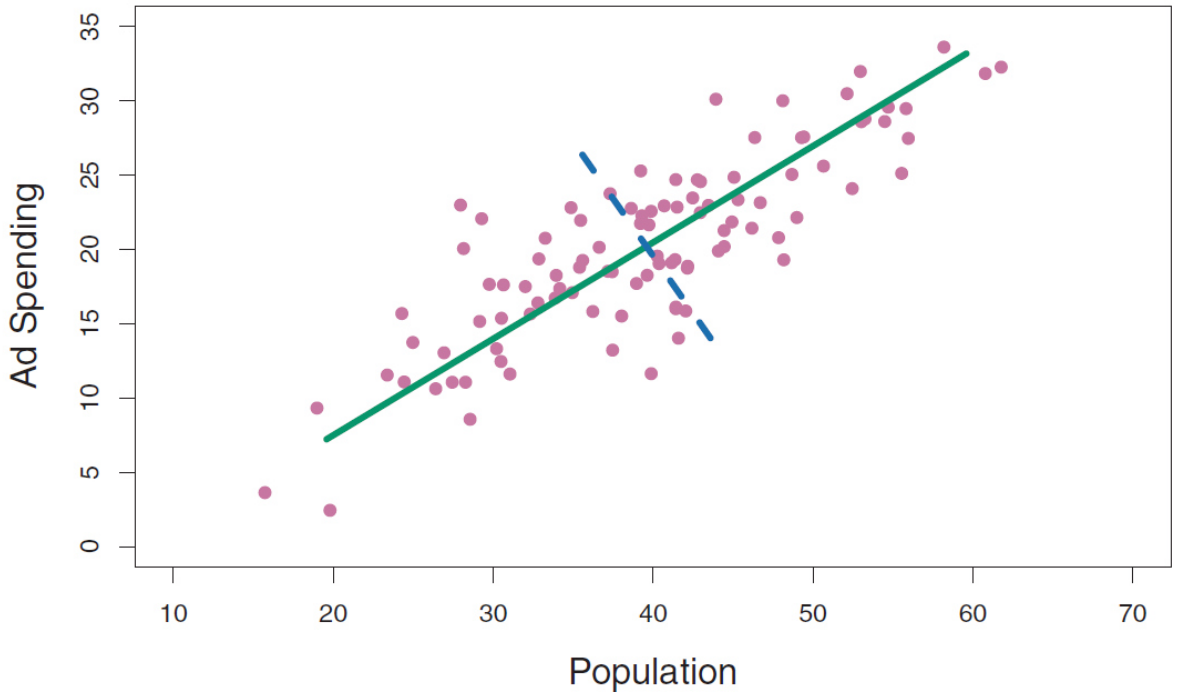

Figure 2.6: Example of the first two principal components (Z_1: green, Z_2: blue) based on the two variables Population and Ad Spending (James et al. 2013, p.240)

The direction of the first principal component Z_1 is along the direction the variables vary the most, i.e. the direction with the biggest and maximal variance. The first principal component can further be seen as the line that is the closest to the original data (Hastie et al. 2009; James et al. 2013). Theoretically, it is possible to compute an infinite number of principal components, however, it makes sense to use only a certain amount of principle components. The second most important one, the second principal component Z_2, is again a linear combination of the input variables. Z_2 however is uncorrelated with Z_1 and has the largest possible variance to the first principle component Z_1 (James et al. 2013).

Choosing the Number of Principal Components

Bro & Smilde (2014) give an overview of different methods to choose the number of principal components. The amount of noise is minimized, the smaller we choose the number of components. The compression of the variation by having less components can therefore lead to statistical benefits in a further statistical modelling process. Each additional component is less interesting than the last one due to smaller value of variation it explains. Dependent on the number of components, also the residuals will change. Accordingly, a reasonable number of components has to be chosen (Bro & Smilde 2014). Although different methods exist, there is no generally applicable one and a combination of several methods should be considered.

The scree test is one of the most common visualization methods to come up with the numbers of principal components (blue line in Figure 2.7). It shows the eigenvalues mapped to its corresponding principal component in descending order. Ideally, the scree plot shows a steep curve with a bending, at which the curve starts to get flat. According to the scree plot, the optimal number of components is where the curve starts to level off (Bro & Smilde 2014).

A second method to determine the number of components is called Kaiser-Guttman's criterion and refers to the number of components that have an eigenvalue higher than one (red line in Figure 2.7). In that case, all the components with an eigenvalue higher than 1 explain more than one of the original variables (Bro & Smilde 2014).

Broken stick is the name of a third alternative which obtains a more realistic cut off for the eigenvalues (Bro & Smilde 2014). An additional line is added to the scree plot that shows how the eigenvalues would be for random data (green line in Figure 2.7). The distribution used for the calculation of this line is called broken stick and symbolizes how the different lengths of a stick would be distributed when broken into different pieces (Bro & Smilde 2014).

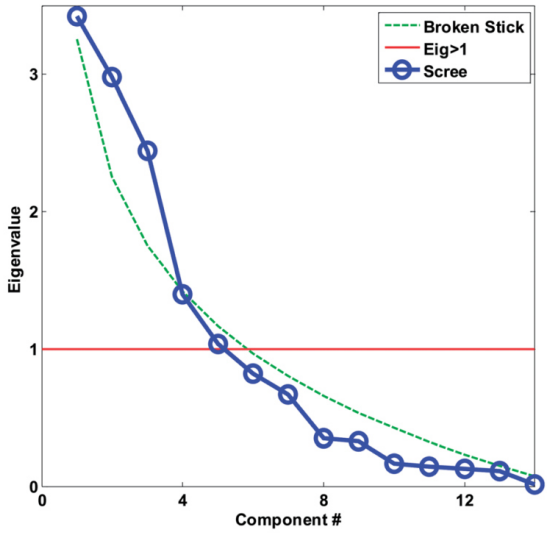

Figure 2.7: Scree plot (blue), broken stick (green) and Kaiser-Guttman's criterion (red) (Bro & Smilde 2014, p.2821)

2.6 Application Areas of Mobile Positioning Data in Human Mobility Research

The in this thesis used data set (a more extensive description follows in **section 3.3**) has some peculiarities. The data is on the one hand based on GPS tuples (latitude/longitude & timestamp) captured by a mobile phone application. On the other hand, the data reflects movement behavior of vehicles, due to the applications purpose as a mobile phone navigation app. Accordingly, we suspect that the data may even give insight into some kinds of touristic patterns.

Based on these peculiarities, the following section is divided into two sections. The first **section 2.6.1** gives an overview of some results of previous research regarding tourist movement. The second **section 2.6.2** then presents examples of studies that have used classification and clustering approaches to study human movement data.

2.6.1 Tourist Movement Research with Mobility Data

Traditionally, research of tourist movements and tourism dynamics are based on surveys and statistics, based on small samples and low granularities (Leng et al. 2016). One of these first traditional studies about spatial movement of tourists was conducted by Fennel (1996) on the Shetland Islands. The underlying goal of Fennel was to better understand how and where tourists move, in order to find out how much pressure tourists exert on the islands. Fennel was, however, aware that the methods used to capture tourist movement admit of improvement. He therefore proposed the adaptation and modification of the radio-telemetry technology that has, at that time, already been used to track animals.

Although Fennel (1996) underlines the need for a better analysis of spatial and temporal behavior of tourists, Shoval & Isaacson (2007) register that only little attention has been paid to the evolution of better methods to analyze tourists in both space and time. Until 2007, the methods used are often limited in accuracy and validity and are most often not directed at the probably biggest aspect of tourism, mobility. Leng et al. (2016) indicate further problems related with such studies: the often unrepresentative sample sizes and the low spatial resolution. However, Shoval & Isaacson (2007) predicted that with the utilization of increasingly more sophisticated mobile phones, it is possible to overcome these problems and to collect more tourist movements.

Subsequently, Shoval & Ahas (2016) remark that the change they prophesized did indeed eventuate. The need for both new data types and data acquisition techniques even has attracted the attention of Eurostat[1], which started several projects to start monitoring tourism with mobile positioning data (Ahas et al. 2014; Shoval & Ahas 2016). Shoval & Ahas (2016) further explain that the first surge of tourist movement studies were dealing with the feasibility of the various methods and tracking technologies for analyzing tourist data. The second surge then deals more with the discovery of spatial phenomena, based on the previously examined methods and technologies.

Furthermore, Shoval & Ahas (2016) emphasize that there are particularly two main approaches to gather the data used for the various tourism studies. With the first method, the researchers must approach tourists actively and present them with a tracking device. An example for this case is for example given by Edwards & Griffin (2013) who tracked 154 participant groups in Sydney and Melbourne to gain better insight into the spatial behavior of tourists in cities. The downside of this approach can especially be seen in the limited number of participants. Their data do, however, have a greater information content, since the participants can be asked about their behavior.

The second passive method is based on data collection from mobile phone networks and micro messaging services (Shoval & Ahas 2016). The downside of this approach is the advantage of the other approach, namely that most of the time no additional information about users are given. It is anticipated however, that due to the mostly huge amount of data (Renso et al. 2012; Zheng et al. 2013), this disadvantage can be compensated.

[1] The statistical office of the European Union (http://ec.europa.eu/eurostat/)

Examples for the second type include, for example, the recent work by Leng et al. (2016) who used CDR data to extract several tourism indicators in the country of Andorra. Among these indicators are flows per country of origin, flows of new tourists, re-visitation patterns, tourist externalities on transport congestion, etc. While some of these indicators are traditionally collected by tourist departments, others could have not been collected without the help of georeferenced data sets, in this case originating from CDR data.

Ahas et al. (2007) analyzed the seasonality of foreign tourist's space consumption in Estonia by using mobile positioning data from anonymized roaming data. They discovered that seasonality plays a significant role in the spatial distribution of tourists, meaning in the case of Estonia that tourists tend to visit the coast in summer and the inlands in winter. Besides the seasonal differences, they further found that tourists of different nations use seasonally different spaces.

A method to combine both spatial data obtained from mobile phones and information from social networks is presented by Girardin et al. (2008). They explored movement of tourists in the city of Rome by looking at mobile phone network data and georeferenced photos. They emphasize that tourists leave two distinct types of footprints, active and passive. User produce active footprints themselves by revealing their locational information in their location tagged photos, text messages and sensor photos. Passive tracks on the other hand arise when tourists interact with the mobile phone network, leading to locational logs and the production of CDR data.

In this thesis, we will build on an observation by Tietbohl et al. (2008), who argue that tourists visiting a new city have a distinct movement pattern. A tourist would visit a museum, go to his hotel, go to a night-club, and then return to the hotel. A stop in his trajectory could therefore be referring to a touristic place, e.g. a hotel or a hotel.

2.6.2 Classification and Clustering with Mobility Data

In the following sections, an overview of studies applying supervised and unsupervised machine learning techniques applied to trajectory mining is presented.

Trajectory Classification

In trajectory classification, class labels of moving objects will be tried to predict based on their trajectories and other features (Lee et al. 2008). It is possible to either classify/cluster the individual objects and their spatio-temporal patterns or directly the individual trajectories itself. In recent years, trajectory classifications has been used for various approaches, including mode of transport detection (Biljecki et al. 2013; Das et al. 2015; Stenneth et al. 2011; Zheng et al. 2008), carpooling profiles (Trasarti et al. 2011) or vessel type detection (Lee et al. 2008).

Several studies have dealt with the development of similarity measures that later can be used to divide distinct groups of trajectories. Pelekis et al. (2012) for example evaluated several similarity measures through a comparison of synthetic and real trajectories. Dodge et al. (2012) on the other hand computed similarity measures between different

trajectory sequences based on movement parameters such as speed, acceleration or direction.

An often-occurring problem in classification approaches is that ground truth is missing to evaluate the results of the classification. Several studies have therefore used alternative approaches to overcome this problem. Renso et al. (2012) conducted an empirical evaluation of the results with domain experts, whereas Pappalardo et al. (2013) compared their findings with traffic counts. Biljecki et al (2013) on the other hand manually classified their trajectories in order to get some sort of ground truth. Another approach is to use contextual information in the trajectories themselves to generate ground truth (Lee et al. 2008).

Categorization of Urban Areas

In recent years, many studies have used mobility data extracted from trajectories to categorize urban areas of different sizes into different groups. These studies can be divided into three categories, based on their underlying objective. Studies of the first category use attractiveness measures to divide areas whereas studies from the second category are dealing with techniques to segment urban areas based on their mobility patterns. The third and last category than deals with land use classification based on mobility data.

An example of the first category is presented by Girardin et al. (2009) who used CDR data to quantify the popularity of an urban area. Using the density and the distribution of aggregate phone calls and photos taken, they came up with a novel way to measure the evolution of attractiveness over time.

Yuan and Raubal's (2012) work is an example for the second category. They used hourly time series of CDR data to measure the dynamic mobility patterns of urban areas. On these time series, they applied a Dynamic Time Warping (DTW) algorithm to measure the similarities between the different urban areas. Similar to that is Reades et al. (2009) who used eigendecomposition on CDR data instead of DTW to identify and extract recurring patterns of mobile phone usage. Based on that, they were able to obtain a higher understanding of the individual places and areas (eigenplaces).

Mobility data can further be used to classify land use, as it is presented by several studies. Pei et al. (2014) used both normalized hourly call volume and total call volume to develop a method to for urban land use classification. They came up with a classification of areas into types such as residential, business, commercial, open space and others. Similar to that is the study by Toole et al. (2012) who used CDR data to measure the spatiotemporal changes in population. Using clustering, they were able to identify groups of areas with similar uses (residential, commercial, industrial, parks and other) and similar mobile phone activity patterns. A third approach of this category is presented by Grauwin et al. (2015) who study the connection between temporal activity profile and land usage in three different cities. Similar to the work by Toole et al. (2012), they use clustering to identify urban areas with similar patterns. Yuan et al.'s (2012) study tries to combine all of the above categories. Using both human mobility data and POI's, they aim to discover regions within a city with different functions.

3. Methodology, Data and Pre-Processing

The following chapter gives an overview of the methodology, data and the pre-processing steps used in this thesis. The first **section 3.1** describes the methodological procedure that ultimately leads to the results of this thesis. In the **section 3.2**, the computing environment is presented. An overview of the movement data used for in thesis is presented in **section 3.3** whereas additional data are presented in **section 3.4**. In **section 3.5**, we introduce the movement behavior ontology that has been used for the database design, presented in **section 3.6**. Finally, **section 3.7** shows an overview of the data cleaning processes that have been applied.

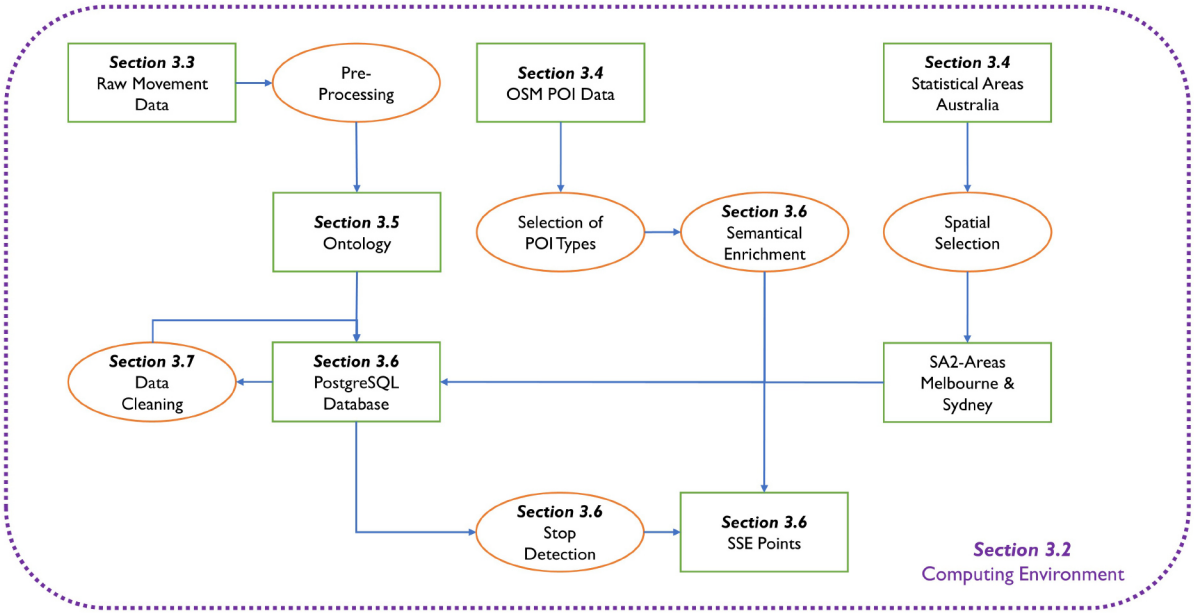

Figure 3.1: Structure and workflow of Chapter 3

3.1 Methodology

This thesis has two main aims. Firstly, we want to categorize users into groups based on their spatio-temporal movement behavior. Secondly, we want to analyze whether the places the found groups visit differ from each other. To get a better understanding of this research aims, an introduction into the methodological procedure as well as some clearing-up definitions are needed.

3.1.1 Definitions

We therefore start with a definition of the *User Type*, which categorizes a user based on its spatio-temporal movement behavior. We define a user type as follows:

User Type: *A user type describes a model of characteristics which are typical for users with a certain spatio-temporal behavior shown in their navigation app usage.*

Based on that definition, we can state that a user type describes the way a user acts, or better, the way he uses the navigation app. Accordingly, a user type can, for example, stand for a pattern that describes a touristic or a commuter behavior. To get to these different user types, we first must form groups of users with similar spatio-temporal movement behavior. We do this by applying a set of (un-)supervised machine learning techniques on the *Spatio-Temporal Footprints* of the individual users. We further define these spatio-temporal footprints as follows:

Spatio-Temporal Footprint: *The spatio-temporal footprint describes the spatio-temporal usage pattern of the navigation app of a single user. Accordingly, it is a proxy for the space usage over time of a certain user.*

The spatio-temporal footprint is a set of measures that describes the spatio-temporal movement behavior and the space usage over time, shown by the way in which the user uses the app to navigate and move around. The spatio-temporal footprint is independent of the actual location the user lives, meaning that the spatio-temporal footprints of two users living at two completely different locations may be similar. After the application of the unsupervised machine learning techniques, we are presented with groups of similar-acting users, revealing similar spatio-temporal footprints. We then describe the spatio-temporal footprints of the individual groups and qualitatively interpret them to come up with the looked-for user types.

In the second research question, we are interested in the differences among the found user types regarding the spatial and the temporal visiting patterns of two cities; Sydney and Melbourne. To address this, we will explore significant locations found along trajectories, called *SSE points*, in each city and for each user type. We therefore define SSE points as follows:

SSE point:	*A SSE point is either a start point, an end point of a trajectory, or the first point of a significant stop segment.*

Thus, a stop segment (defined in **section 3.6.2**) can be seen as a significant stop along the route of the trajectory and can be operationalized based on various types of thresholds (defined in **section 3.6.3**). The SSE points stand for the actual places the individual users have visited. The location of these places is independent of the spatio-temporal footprints of the users and can therefore be used to analyze their patterns. The aggregate of all SSE points per user type will then be analyzed for different areas of the cities to get a deeper understanding of both the temporal and spatial characteristics of the different user types.

3.1.2 Methodological Procedure

The structure of this thesis follows the methodological procedure as presented in Figure 3.2. We start with a characterization of the data used in this thesis, first the movement data provided by Sygic in **section 3.3**, followed by a description of the additional data in **section 3.4**.

For the design of the database that stores both the Sygic as well as the additional data, we use a movement behavior ontology that is presented in **section 3.5**. The actual database design and its implementation is then outlined in **section 3.6**. A lot of data cleaning was needed to reduce the influence of outliers. We have therefore set various standards to remove flawed points, trajectories and users which will be presented in **section 3.7**. In that section, we further present our approach to segment the trajectories, involving the stop detection in **section 3.6.3**. The detection of stops leads us to the formation of the SSE points (**section 3.6.4**) that will later be used to analyze the spatial and temporal differences of the found user types in Melbourne and Sydney.

For all users remaining after the data cleaning, we have computed a set of measures that describe their spatio-temporal footprints (**section 4.1**) We then have taken only users that have used the app while at least 5 days, and have applied a Principal Component Analysis (PCA, **section 4.2**) on their spatio-temporal footprints. We then have applied different clustering methods on various numbers of principal components to find the best method separating users into distinct groups (**section 4.3**). In **section 4.5**, the found clusters have then been qualitatively described and interpreted to form the different user types.

In a next step, the SSE points of the individual users of the established user types have been aggregated to examine their temporal and spatial characteristics in the two cities of Melbourne and Sydney. Therefore, only users have been chosen that have visited either of the two cities on at least one day. The chosen methods to analyze the differences in the temporal and the spatial characteristics of the user types are described in **sections 5.1** and **5.2.**

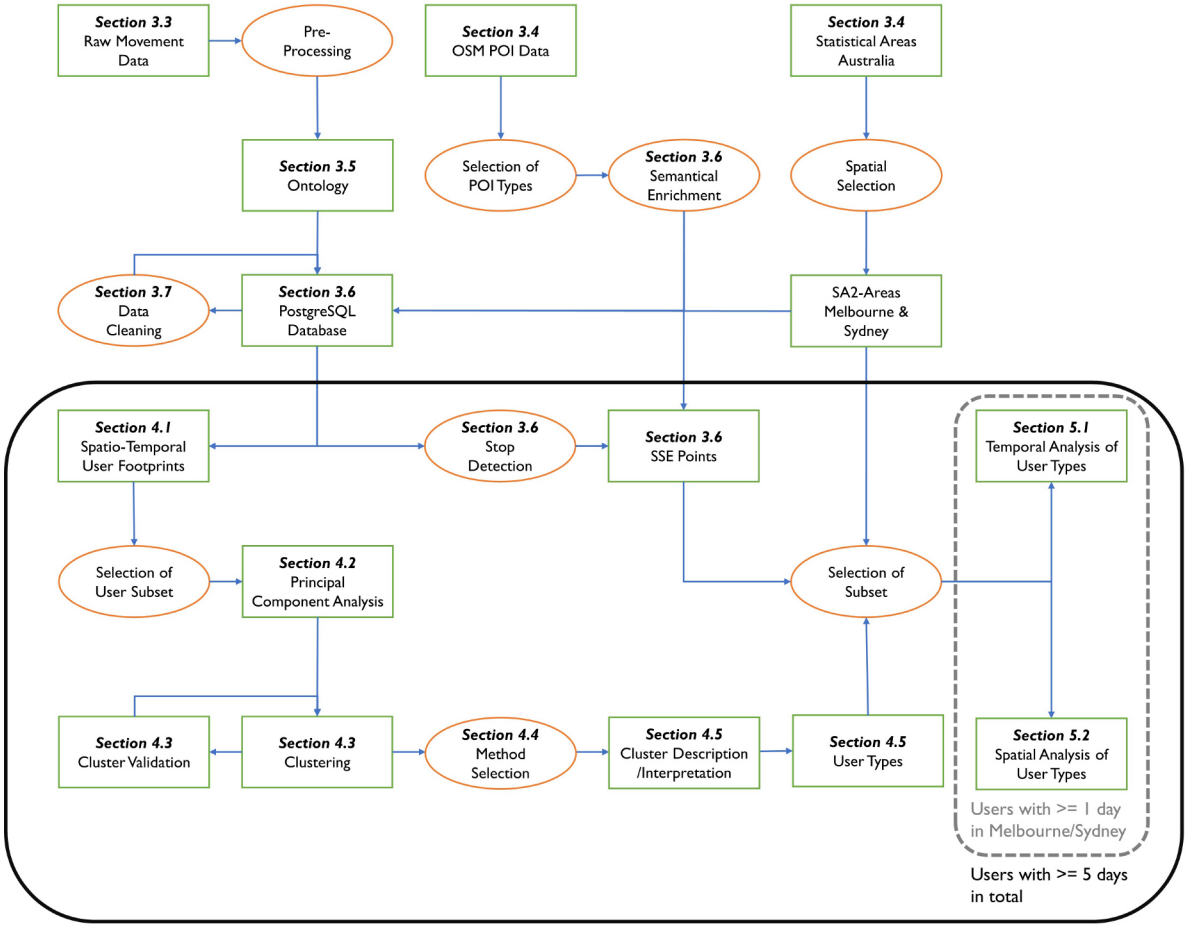

Figure 3.2: Methodological framework of this thesis

3.2 Computing Environment

For the hosting of the database, a Linux server (Ubuntu 16.04) with 400GB memory and 16GB RAM was set up. To store the movement data used for this thesis, PostgreSQL 9.5[2], an object-relational database was chosen. Due to the spatial characteristics of the data, additionally PostGIS 2.2[3] and the PostGIS topology extension were used. PostGIS is a spatial database extension for PostgreSQL and enables the storing of geometry data types such as points, lines, and polygons. Additionally, R (R Core Team 2016) and RStudio-Server[4] (version 0.99.902) were installed on the server and used as the analytical environment.

Most data management and processing steps were made in PostgreSQL (**section 3.6** to **section 4.1**). The computation of all statistics (from **section 4.2** onward) as well as their visualization were done in R. Several R-packages were used in this thesis, whereas the most important ones are listed in Table 3.1.

[2] https://www.postgresql.org/
[3] http://www.postgis.net/
[4] https://www.rstudio.com/products/rstudio-server/

Table 3.1: An overview of the most important R-packages used in this thesis

R-Package Name	Task	Reference
RPostgreSQL	Interface to PostgreSQL-Database	Conway et al. (2016)
dplyr	Data Manipulation	Wickham & Francois (2016)
tidyr	Data Manipulation	Wickham (2016)
lubridate	Temporal Data Manipulations	Grolemund & Wickham (2011)
rgeos	Spatial Data Manipulations	Bivand & Rundel (2016)
factoextra	PCA Visualizations	Kassambara & Mundt (2016)
cluster	Clustering Methods	Maechler et al. (2016)
clValid	Cluster Validation	Brock et al. (2008)
RankAggreg	Rank Aggregation	Pihur et al. (2009)
spdep	Spatial Autocorrelation	Bivand & Piras (2015)
ggplot2	Statistical Visualizations	Wickham (2009)
sp	Spatial Objects in R & Choropleth Maps	Bivand et al. (2013)
leaflet	Interactive, Map-based Visualizations	Cheng & Xie (2016)
RColorBrewer	Colors used for Visualizations	Neuwirth (2014)

3.3 Sygic Data Characteristics

The in this thesis used data originates from the navigation app producer Sygic[5]. Sygic produces a wide range of mobile phone navigation apps for both Android and iOS, including *GPS Navigation, Car Navigation, Speed Cameras, Truck Navigation, Taxi Navigation* and *Travel*. Sygic's apps have over 150 million users in total, which makes it the navigation app producer with the second highest user count (Sygic 2016). To reduce the grade of confusion, the individual app types will be ignored in the remainder of this thesis and, accordingly, only the term "app" will be used.

The data provided by Sygic consists of GPS tuples (latitude/longitude position & timestamp) which are enriched with additional information. Each tuple possesses a unique ID of both the user and the session it belongs to. All the attributes, including the ones further being used for this thesis (highlighted) can be seen in Table 3.2. We further define a session as follows:

Session: *A session is a set of timestamped recordings, including GPS locations, of a person using the app to navigate.*

[5] http://www.sygic.com/gps-navigation

Table 3.2: Attributes of the original data, as delivered by Sygic. Highlighted in gray are the further used attributes.

Variable Name	Description
sensortime	The timestamp (date and time) of the GPS tuple.
latitude	The latitude of the GPS tuple.
longitude	The longitude of the GPS tuple.
heading	Heading value in degrees (0 to 359) at a GPS tuple; if heading is invalid or not available it equals -1
speed	The speed at a GPS tuple, computed directly by the device.
altitude	The altitude of the device at a GPS tuple.
haccuracy	
vaccuracy	
computedbearing	
computedspeed	
foreground	
networktype	The network type at the time the app is being used [0, 1, 2, 3]
regioncode	
sessionid	The unique ID of a session.
deviceid	The unique ID of a device-
platform	
devicemodel	
osversion	
advertisingid	

3.3.1 User Identification

A unique device ID (*deviceid*) has been assigned to each device on which a Sygic navigation app is installed. Hence, it is not possible to identify a unique user, but a unique device. Accordingly, a person might use the app on several devices or, several people might use the app on the same device. To speak of a unique user in that case is therefore not entirely correct. For the sake of convenience however, an individual user is defined as the user of a single device. For the remainder of this thesis, the term user ID will therefore be used.

3.3.2 Trajectory Identification

According to the app producer, each time a user issues a navigation query and starts the navigation, a new session ID (*sessionid*) is generated. Unless the user terminates the navigation, a session keeps being recorded. Accordingly, no new session is generated when the app is put on standby. When a user enters a tunnel or the loses coverage, however, the data are not cached.

Based on some investigations into the data's structure, we have seen that some sessions start almost after the finishing of the previous session. For computational reasons, we

have ignored this, and therefore see a session as a trajectory leading from a start point to an end point.

3.3.3 Temporal and Spatial Resolution

According to Sygic, the individual GPS tuples provided are not map-matched and are provided in their raw form, as acquired by the sensors in the user's mobile devices. The data further come along with timestamps in coordinated universal time (UTC). The stated temporal resolution is five seconds, although this varies from three to nine seconds.

3.3.4 Spatial and Temporal Extent

For the spatial extent, we have chosen the country of Australia. Australia (Figure 3.4) as a country is an interesting study region to its isolation and associated with it, its lack of direct border crossings from/to other countries. To access Australia, almost all tourists need to border the country via plane and are dependent on rental cars or public transport when moving around in Australia. Moreover, Australia has a high level of urbanization (The World Bank 2015) and an increasing amount of traffic (Department of Infrastructure and Transport 2012), which makes it an ideal testing ground.

Figure 3.3: Map of the spatial extent of the data used in this thesis, Australia. Source: Maps of World (2013)

After considering several statistics regarding tourist movement in Australia, the month of January 2016 was chosen as the period of investigation. As Figure 3.3 shows, January 2016 has had about average oversea tourists when compared to previous and following months (Australian Bureau of Statistics 2016b). January 2017 additionally shows high amounts of domestic overnight tourists in Victoria and New South Wales (Tourism Research Australia 2016). Besides that, January is the month where several major events are held throughout Australia, such as the Australian Open, one of the most visited tennis tournaments worldwide.

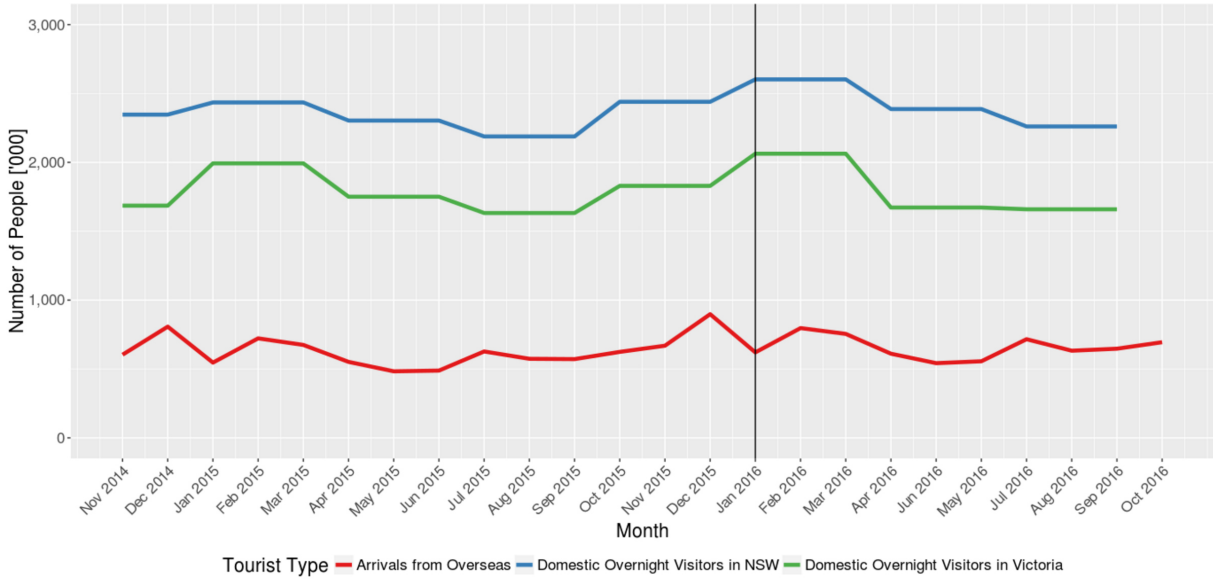

Figure 3.4: Number of short-term international arrivals, domestic overnight visitors in New South Wales (NSW) and Victoria per month (Australian Bureau of Statistics 2016b; Tourism Research Australia 2016)

3.3.5 Peculiarities of the Data

Several things are special about the data used for this study, when comparing to other known data sets from previous studies. First, the data does have a very high spatio-temporal resolution (five second interval) for the time the app has being used. Accordingly, it allows us to have a very detailed look at the individual trajectories. A disadvantage of that is, that we do not have additional data for the time the app is not being used, i.e. the user is moving around without the app running as navigation aid.

A second peculiarity is the specific purpose of the app itself: navigation. Accordingly, we hypothesize that users only use the app in case they want or need additional spatial information about their route. This spatial information may include information about the route itself or information about the road conditions, congestion, and speed cameras. Based on that, we do not have continuous information whenever a certain car is used, but only whenever additional information is needed by said user. The data therefore presents a high spatio-temporal resolution, however, only when the app is really used and not always a user drives his car.

This stands in contrast to other data used for mobility studies, such as the data used by Pappalardo et al. (2013), which is originating from on-board GPS receivers from cars. Such data was collected whenever the car was used, regardless of the task and back-

ground of the drive. On the one hand, this leads to an even higher stream of data as the one we are presented in this thesis. On the other hand, however, the data might originate only from the local population, which indirectly leads to the third peculiarity of our data. The data used in this thesis comes from a mobile app and is therefore acquired by the smartphone itself. That said, the navigation app may be used by several types of users, including professional drivers who regularly use smartphones for navigating.

The app's main purpose is to help navigating. We therefore assume that most of the users are not familiar with their surrounding in which they are driving in. Moreover, we argue that a certain proportion of users can be considered as non-native or even touristic. The proportion of tourists/locals can therefore be considered higher as in other data sets used for human mobility research.

3.4 Additional Data

Besides the data provided by Sygic, we are additionally using two other types of data, a demographic data set based on the census of the Australian population by the Australian Bureau of Statistics, aggregated by Statistical areas; and OSM (OpenStreetMap) map features.

3.4.1 Census Demographic Area

The two most populated Australian cities, Sydney, and Melbourne, were chosen to analyze the spatial characteristic of the SSE points. In 2011, the Australian Bureau of Statistics created a new framework to divide the country into different statistical areas, called Australian Statistical Geography Standard (ASGS; Australian Bureau of Statistics 2016a). Of the five different main structures of ASGS, two are of interest for this thesis, namely the Statistical Areas Level 1 (SA1) and the Statistical Areas Level 2 (SA2). According to the Australian Bureau of statistics (2016a), SA1 units are the smallest statistical unit for richer demographic data are released. Approximately 55'000 SA1 units cover the whole of Australia, whereas in each unit, there is an approximate population of 400 people. SA1 areas aggregate to larger SA2 areas with each about 10'000 inhabitants. For the whole of Australia, about 2'196 SA2 units exist (Australian Bureau of Statistics 2016a).

For this thesis, both SA1 and SA2 data for the metropolitan areas of Melbourne and Sydney were chosen. A first analysis showed that the SA2 areas offer more comprehensive data than SA1 areas, at a level necessary for this thesis. Due to that, only SA2 areas were further used in the analysis.

Table 3.3: Background information on the two cities (Australian Bureau of Statistics 2015)

City	Area (km²)	Population	Population Density (people/km²)
Melbourne	9'138.75	~4'529'500	453
Sydney	10'687.12	~4'921'000	400

3.4.2 OSM Map Features

OSM data was used to semantically enrich the start, stop and end points (SSE points). Data was extracted from OSM via its API, Overpass Turbo[6], whereas a focus was set on cartographical elements that could be important for a navigation app user. We defined these places as POI's (points of interest) for which we expect users to look for when using the app. The respective elements (see Table 3.4) were selected after an examination of the different map features of OSM[7].

OSM data can consists of three different types: nodes, ways, and relations. Nodes define points, whereas ways can both define linear features and boundaries (lines and polygons). Since relations can be consists of several nodes and ways at once, only nodes and ways were used for this thesis. The downloaded nodes and ways were then stored in the database.

Table 3.4: Overview of the downloaded OSM map-features as stored in the database (state: January 23rd, 2017)

OSM Key	OSM Value	# of POI's	# of Lines	# of Polygons
tourism	attraction	1'523	12	420
tourism	viewpoint	4'447		
tourism	hotel	1'139		733
tourism	hostel	239		71
tourism	motel	1'240	1	563
tourism	guesthouse	365		78
tourism	camp_site	5'740		557
tourism	caravan_site	602		1'223
tourism	chalet	138		187
tourism	theme_park	22		31
tourism	zoo	24		51
tourism	museum	608		235
amenity	car_rental	186	1	29
amenity	restaurant	5'088	1	636
amenity	fast_food	4'371	2	629
amenity	bar	617		48
amenity	pub	2'809	1	792
amenity	cinema	221		62
amenity	nightclub	80		22
amenity	theatre	151		113
amenity	marketplace	38		46
amenity	place_of_worship	2'855		1'784

[6] https://overpass-turbo.eu/
[7] http://wiki.openstreetmap.org/wiki/Map_Features

amenity	parking	6'807		31'671
amenity	cafe	4'635		476
amenity	kindergarten	458	1	380
amenity	school	1'033	5	7'391
amenity	college	70	2	334
amenity	university	33		246
amenity	hospital	352		710
shop	mall	56		870
shop	supermarket	2'797		634
leisure	fitness_centre	98		15
leisure	sports_centre	637		1'470
leisure	stadium	16		180
public_transport	station	1'071	12	135
office	*	1'129		393

3.5 Mobility Behavior Ontology used for Database Design

We are adapting the ontology of mobility behavior from Renso et al. (2012) for our database design. They propose a mobility behavior ontology based on two conceptual levels (Figure 3.5). The core ontology (orange) describes the concept of human behavior, independent of a specific application domain. It uses the concepts of trajectory, stop, move, time and pattern. The application ontology in green and blue consists of various elements that are relating to the application context. In the given case study designed by Renso et al. (2012), the application ontology relates to the movement of tourists in urban spaces, e.g. visiting tourist places or staying at a particular accommodation.

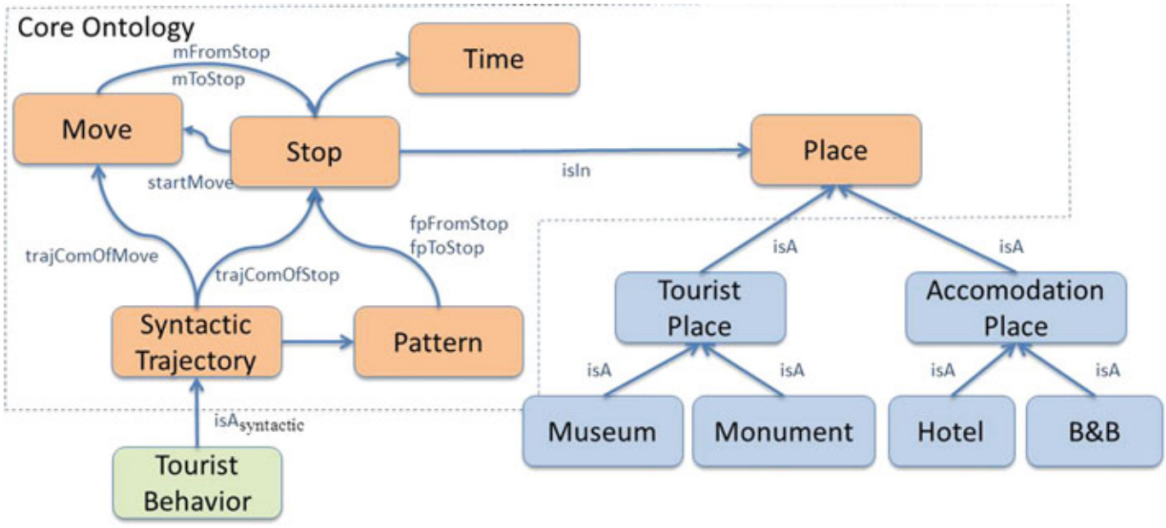

Figure 3.5: An example of a possible mobility behavior ontology defined by Renso et al. (2012, p.34). Core ontology elements are emphasized with orange boxes, application ontology elements in blue and green.

3.6 Conceptual Database Design (ER-Model)

Based on the possibility of identifying both user and trajectory and the ontology in Figure 3.5, a first draft of an entity-relation model (ER-Model) was established (Figure 3.7). This first draft consists of three entities whose information can be found in the raw movement data. The first entity, called *user,* stores information about the individual users whereas the *session* entity stores information about the individual sessions. The *point* entity stores information about the individual tuples (coordinates & timestamp).

As shown in Figure 3.7, each user can generate one or more sessions which then consists of one or more points. Each session can, however, only be assigned to one user. This is similar with the points, which each can only be assigned to one session. Both the entities *user* and *session* have their respective ID's already through the original data, the primary key of the *point* entity, however, had to be established first.

3.6.1 Adding Movement Information

The session entity stores information about trajectories. To define a trajectory, we adapt the definition of Renso et al. (2012) as follows:

Trajectory: *A trajectory is the footprint of different positions of a moving object. It is a sequence of tuples recorded by a tracking device* (adapted from Renso et al. (2012)).

Accordingly, lining up all points of a session then forms a trajectory. The trajectories found in the data show the individual users spatio-temporal behavior over a specific amount of time. If we sum up the individual trajectory over an overlying time period, we come up with movement tracks (Figure 3.6), as suggested by Parent et al. (2013). These movement tracks can be defined as follows:

Movement Track: *A movement track is the sum of all recorded positions of an individual over a defined period* (adapted from Parent et al. (2013)).

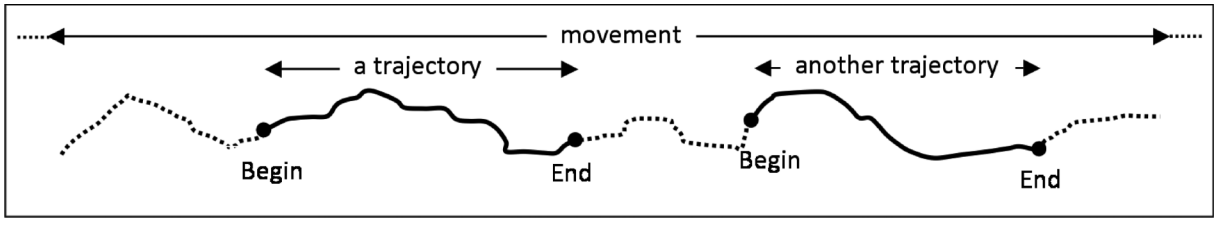

Figure 3.6: Trajectories (bold lines), movement tracks (dotted and bold lines) and the whole movement (Parent et al. 2013, p.4)

Based on that, a new entity storing information about a users' daily activity was created, named *dailymovement* (Figure 3.8). *dailymovement* is related to a movement track over the period of a day and describes the sum of all trajectories over that period.

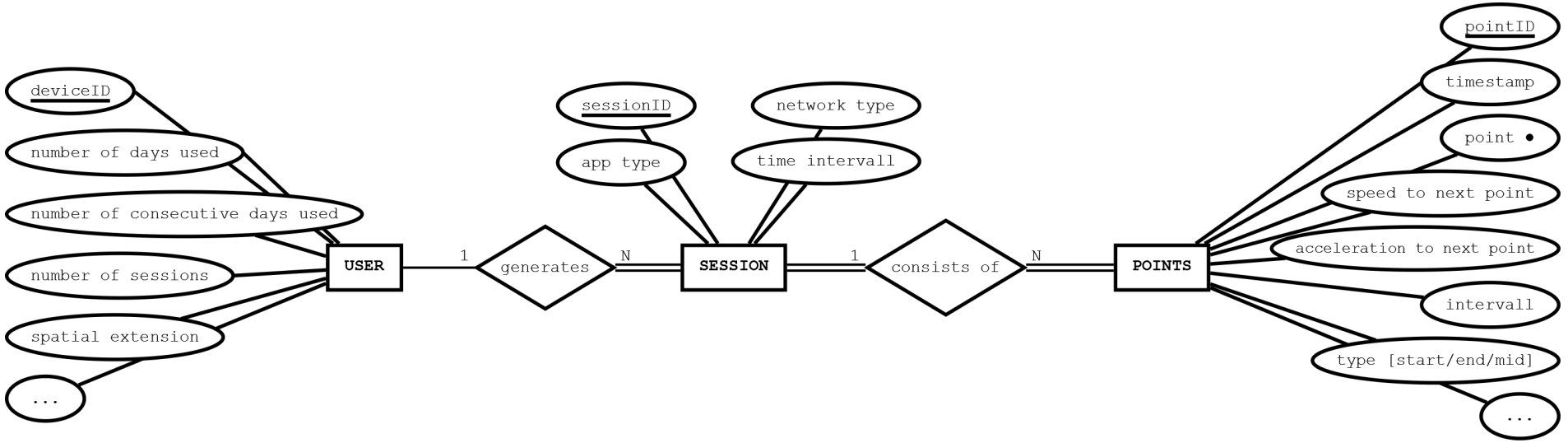

Figure 3.7: First draft of the ER-Model

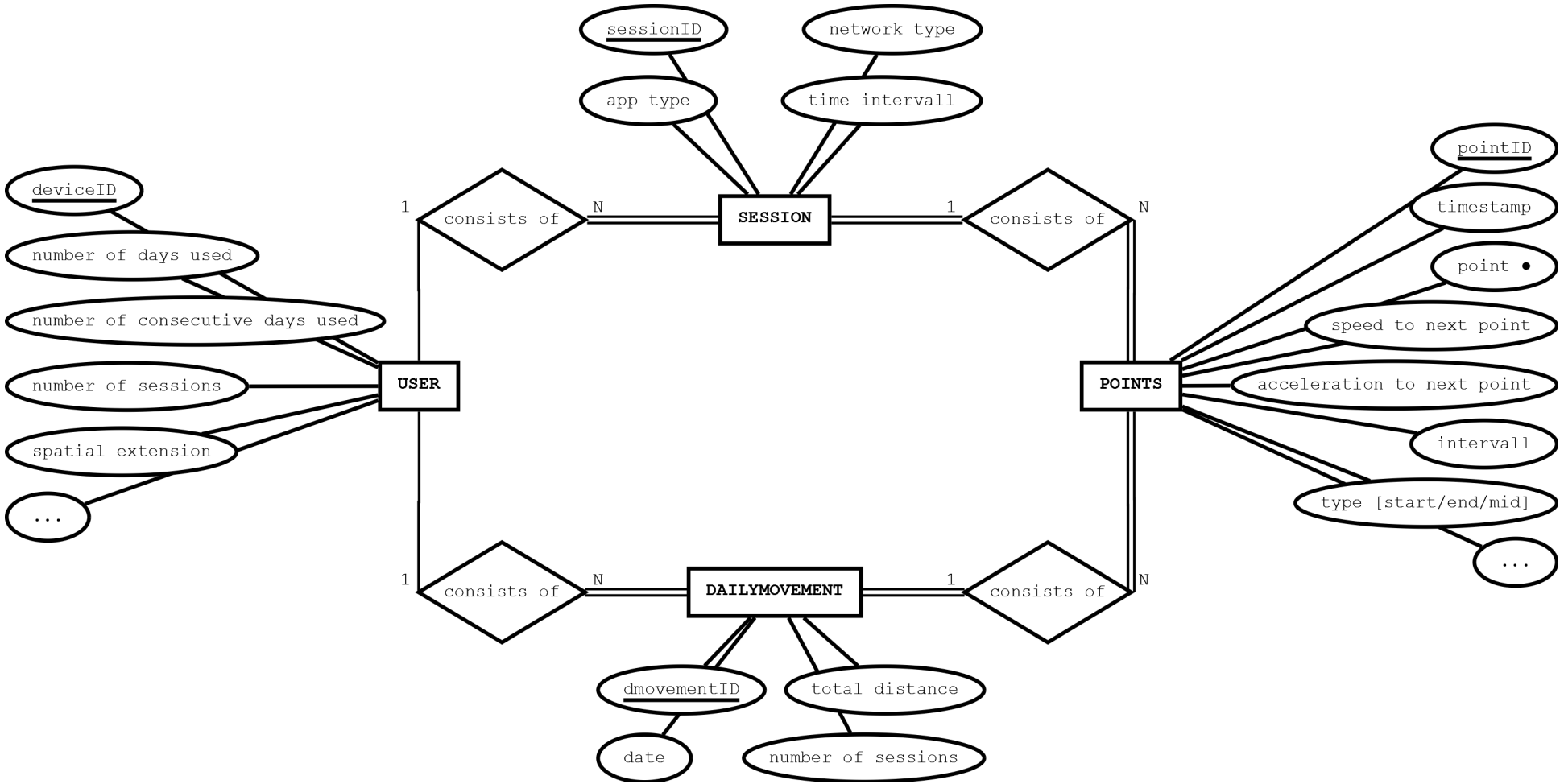

Figure 3.8: Second draft of the ER-Model

3.6.2 Semantic Classification of Trajectory Segments

Although most trajectories show some movement behavior, we cannot assume that a user was continuously moving around while producing a trajectory. As Spaccapietra et al. (2008) emphasize, trajectories may themselves be semantically segmented into different time intervals. In this thesis, we have classified each point of a trajectory into *start*, *stop*, *move* and *end*, as it is proposed by Spaccapietra et al. (2008). Accordingly, a trajectory is a sequence of *moves* from a *start* to an *end* point, intersected by *stops*. We therefore define the terms *start point*, *end point*, *stop segment* and *move segment* as follows:

Start Point: *A start point is the first tuple of a trajectory.*

End Point: *An end point is the last tuple of a trajectory.*

Stop Segment: *A stop segment is a sequence of tuples where the distance between any adjacent position is less than a spatial threshold and the time spent within the sequence is greater than a time threshold* (adapted from Phithakkitnukoon et al. 2010).

Move Segment: *A move segment describes a segment of a trajectory which shows continuous movement and no stop segments. Accordingly, it starts either at a start point or the last point of a stop segment and ends at the last point of a stop segment or at end point of a trajectory.*

Furthermore, we define the summed up and covered distance in a move segment as follows:

Step Length: *The step length is the sum of covered distance between two tuples in a move segment. It must not be mistaken with the direct Euclidean distance.*

The overlying goal of the approach to classify segments and individual points of the trajectory is to obtain so called *semantic trajectories*. This can be achieved by attaching semantics to the stop segments as well as to the start and end points of the individual trajectories.

Semantic Trajectory: *A semantic trajectory is a trajectory that has been enhanced with annotations regarding the individual segments* (adapted from Spaccapietra & Parent (2011) and Parent et al. (2013)).

3.6.3 Stop Detection

In the last section, a stop segment has been defined. A literature review showed that several methods can be carried out to compute the individual stop segments in trajectories. Zheng et al. (2011) and Andrienko et al. (2013) computed stops as sequences of tuples whose spatial and temporal extent is below a certain threshold. A similar approach is presented by Alvares et al. (2007) with their SMoT-method. Here, a stop is defined as a position where a trajectory stops for a certain amount of time at a, by the application predefined, POI. A third approach is presented by Tietbohl et al. (2008), who propose a speed-based spatio-temporal clustering approach to detect stop segments.

Based on a thorough data examination, we decided to compute the stop segments based on spatio-temporal thresholds, similar to Andrienko et al.'s (2013) approach. The goal of our stop detection approach was to find stops that that enable a user to execute a minimal task. We have therefore tested several temporal and spatial thresholds on a small set of trajectories and have further visualized the resulting trajectories and the found stop segments, respectively. The resulting visualizations then have shown that a temporal threshold of five minutes and a spatial threshold of ten meters generate reasonable stop segments that stand in great contrast to stops on a smaller temporal level, that could arise due to congestion. Within a stop lasting at least five minutes, the respective user can interact with the environment and the actual location. Accordingly, we have come up with the following working definition of a stop point:

Stop point:	*A stop point is the first tuple of a stop segment. A stop segment is declared as a series of tuples, with almost no movement (> 10 meters) in the next 5 minutes.*

To compute the sum of the covered distances over the next 5 minutes for each tuple, a moving window was applied on each entry of the point entity. The corresponding SQL-query for that can be seen in the Code Fragment 1 in Appendix .

3.6.4 Implementation of SSE Points

In the given case of this thesis, it is not entirely clear whether the app stops the recording of a session when being in stand-by for too long. Accordingly, it is possible that stop events are manifested in both the actual stop points and the end points of the trajectories. Based on that, we decided to store the start, stop and end points all together in one single entity.

To store information about both the move segments and the SSE points, we came up with new entities called *move* and *start_stop_end_point* (see Figure 3.9). The *move* entity stores information such as the step length (the distance covered in a move segment) as well as the average speed in the move segment.

The entity *start_stop_end_point* consists of the start and end points of the individual trajectories as well as the first point of the stop segments, as defined above. The SSE points were further semantically enriched with OSM POI's.

3.7 Data Cleaning

Table 3.5 gives an overview of parameters set for considering the recorded trajectories and users in the analysis. This is a data cleaning process aiming to reduce the influence of outliers impacted by data collection errors on the analysis. The therefore applied operations are not only parameters, but more a sequence of filtering operations that have been applied in the specified order.

Table 3.5 further shows that the relative amount of data lost is lower when looking at the sole number of tuples than when looking at the number of sessions. Accordingly, most of the discarded sessions consisted of small amounts of GPS tuples.

Table 3.5: Self-set data standards and its effect on the amount of data

Operation	No. of Tuples remaining	No. of Sessions remaining
Original Data	166'200'862	
Removal of all points with no session ID, device ID and coordinates or wrong timestamp	160'030'885	2'363'140
Removal of all sessions with less than 24 points (less than 2 minutes long)	147'328'974	
Removal of all sessions that are shorter than 300 meters	127'611'519	545'566
Removal of all sessions that have velocities higher than 50 m/s (180 km/h)	127'047'175	545'553
Removal of all users with less than 10 distinct points	126'872'144	545'444
Removal of users whose number of sessions is much smaller than the number of days used	126'468'040	545'440
Relative amount of data remaining	76.09%	23.08%

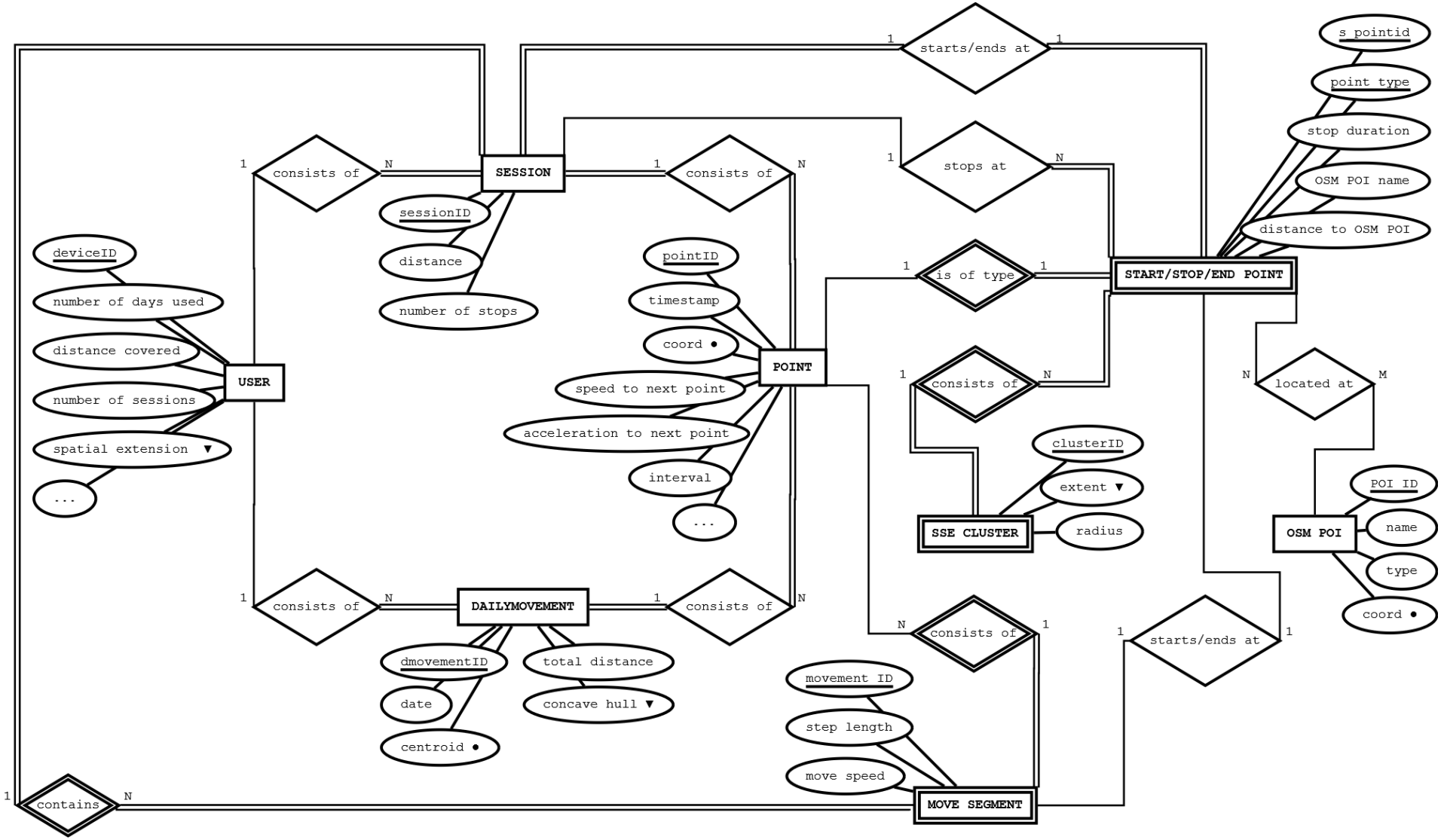

Figure 3.9: The complete ER-Model of the PostgreSQL-database

4. From Users to User Types

The following chapter describes the data analysis methods used that ultimately lead to the worked-out user types (seen in Figure 4.1). Presented from **section 4.1** onwards is the workflow of the categorization of the individual users into different user types. This process starts with **section 4.1** showing the different measures that have been calculated for each user, describing each user's spatio-temporal footprint. In **section 4.2**, we describe the application of PCA as a means to reduce the dimensionality of the data. Following that is **section 4.3**, in which we study the characteristics of dominant, cohesive clusters of users based on the most significant principal components identified by the PCA. In **section 4.4**, a definite clustering approach is chosen based on the results of the clustering validation in **section 4.3.1**. Finally, in **section 4.5**, we describe the individual clusters and interpret quantitatively, leading up to the formation of user types.

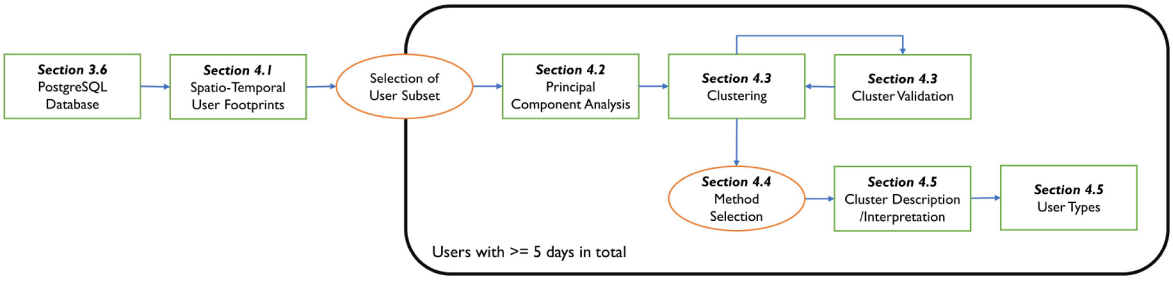

Figure 4.1: Workflow of Chapter 4

4.1 User Measures – Spatio-Temporal Footprints

To characterize users into different cluster that later can be interpreted to form user types, 36 measures[8] were calculated that reflect the users spatio-temporal footprints (Table 4.1). From these 36 measures, the first 33 measures (highlighted) are further

[8] In machine learning, such measures are often called features or attributes.

used for the PCA, whereas the other three measures are only used for the qualitative description and interpretation of the later found user types in **section 4.5**.

These measures were computed based on spatial, temporal, and spatio-temporal characteristics of the different users. The goal was to establish a spatio-temporal profile, i.e. footprint for each user, which can be used for a best possible division of the individual users into distinct groups. No aggregate measures about sessions were computed due to problems with session characterization. Due to that, only daily movement pattern characteristics were computed that do not suffer as much from the uncertainties in the session definition.

Table 4.1: Overview of the calculated measures per user. Shaded measures are used for the PCA whereas non-shaded measures are only used for qualitative interpretation of found user types.

	Variable name	Clarification and Description
1	num_days	The number of days the user has used the app.
2	num_cons_days	The highest number of consecutive days the user has used the app.
3	num_dstnct_wdays	The number of distinct weekdays the user has used the app.
4	period	The number of days between the first and last usage.
5	tot_dist	The total distance the user has covered [m].
6	tot_time	The total amount of time the app has been running [s].
7	area	The area of the concave hull of all sessions [m^2].
8	circum_hull	The circumference of the concave hull of all sessions [m].
9	max_dist	The maximum distance within the concave hull of all sessions [m].
10	complexity	The complexity of the concave hull (area/circumference).
11	compactness	The compactness of the concave hull [$4*area/\pi*max\_dist^2$].
12	mean_d_dist	The average distance covered in a day [m].
13	sd_d_dist	The standard deviation of the average distances per day [m].
14	mean_d_area	The average area of the daily concave hulls [m^2].
15	sd_d_area	The standard deviation of the daily concave hulls [m^2].
16	mean_d_overlp_pc	The average percent of overlap of two consecutive[9] daily concave hulls [%].
17	sd_d_overlp_pc	The standard deviation of the percentage of overlap of two consecutive daily concave hulls [%].
18	mean_d_cent_dist	The average distance of two consecutive daily centroids [m]. The centroid is the centroid of the daily concave hull.
19	sd_d_cent_dist	The standard deviation of the distance between two consecutive daily centroids [m].
20	mean_dist_overall_cent	The average distance between the daily centroid and the overall centroid [m].

[9] in this and the following cases as well, *consecutive* refers to the next day the app has been used, which does not necessarily mean the day after.

21	sd_dist_overall_cent	The standard deviation of the distance between the daily centroid and the overall centroid [m]
22	num_move	The absolute number of move segments in all sessions.
23	mean_step_length	The average distance covered in a move segment [m].
24	sd_step_length	The standard deviation of the step length [m].
25	mean_move_speed	The average speed in the move segments [m/s].
26	sd_move_speed	The standard deviation of the speed in the move segments [m/s].
27	num_stops	The number of stops in all sessions.
28	sse_osm	The number of SSE points near OSM POIs.
29	tot_stop_dur	The total duration of all stops [s].
30	mean_stop_dur	The average duration of a stop [s].
31	sd_stop_dur	The standard deviation of the stops [s].
32	sd_time	The standard deviation of all time stamps [s].
33	num_clusters	The number of clusters of SSE points.
34	days_melb	The number of days the user has spent in Melbourne.
35	days_syd	The number of days the user has spent in Sydney.
36	days_scen	The number of days on one of the six scenic roads.

4.1.1 Temporal Measures

We both considered absolute temporal measures such as the absolute number of days (*num_days*), as well as relative temporal usage measures such as the highest number of consecutive usage days (*num_cons_days*). Besides that, we computed the number of distinct weekdays (*num_dstnct_wdays*), giving us an overview on how regularly the app is used in terms of different weekdays. Another temporal measure is called *period* and describes the number of days that lie between the first and the last usage in the month of January.

4.1.2 Daily Area and Centroid

An area per day is first needed to later compute the daily overlap of two consecutive days. The area can be computed based on several different assumptions and functions such as, for example, the convex hull or the concave hull. Both geometries represent a geometry that encloses all given geometries, i.e. points. As it can be seen in Figure 4.2, concave hull encloses the geometries in a far better way. PostGIS (2016) describe the concave hull as the "*geometry you get by vacuum sealing a set of geometries*", whereby the smaller the chosen target percent, the smaller the area in comparison to a complex hull.

For the computation of the concave hull and the centroid of the concave hull, the two PostGIS-functions *ST_ConcaveHull* and *ST_Centroid* were used (PostGIS 2016). An example of the application of that function can be found in the Code Fragment 4 in Appendix C.

Based on the concave hull, several user values were then calculated, including the overall area (*area*), the mean daily area (*mean_d_area*), the standard deviation of the daily area (*sd_d_area*), the mean overlap of the areas of two consecutive days (*mean_d_overlp_pc*) as well as its standard deviation (*sd_d_overlp_pc*). Furthermore, the circumference of the overall area (*circum_hull*) as well as the complexity and the compactness were calculated based on the concave hull function.

The daily centroid was further being used for the calculation of the mean distance between two consecutive daily centroids (*mean_d_cent_dist*) and its standard deviation (*sd_d_cent_dist*). Both the daily and the overall centroid were then used for the calculation of the mean distance between the overall and the daily centroid (*mean_dist_overall_cent*) and its standard deviation (*sd_dist_overall_cent*).

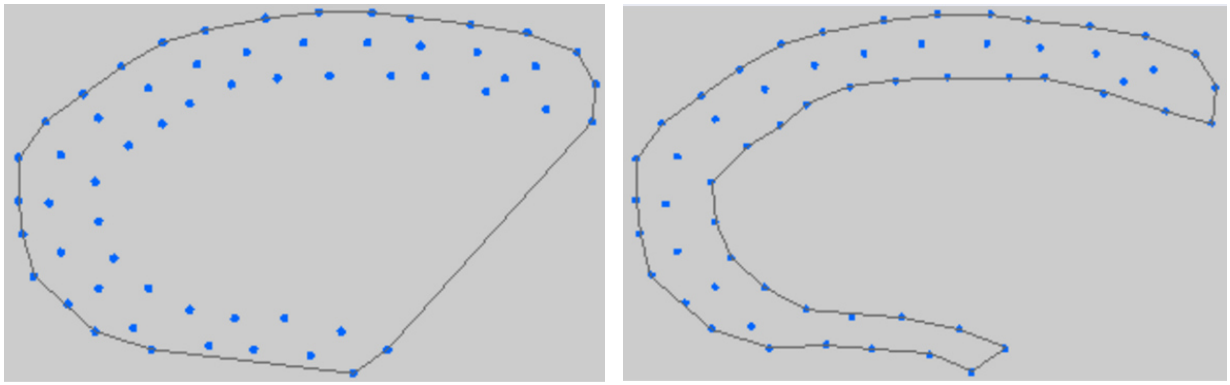

Figure 4.2: The convex hull (left) and the concave hull (right) for the same set of points (Ubicomp@UMinho 2006)

4.1.3 Number of Spatial Clusters

The number of spatial clusters for each individual user was calculated using the *ST_ClusterWithin*-function, due to the unavailability of the new *ST_ClusterDBScan*-function in version 2.2 (see Code Fragment 5 in Appendix C). Like the DBScan-algorithm, ST_ClusterWithin takes an *eps* value as input, representing the minimum distance from which the individual points must be separated from each other to become part of a cluster. The ST_ClusterWithin-function does, however, not have a *minpoints*-input parameter. Accordingly, already single points can form a cluster.

We tested several eps-values for their usability and finally came up with 15 meters as an appropriate eps-value. This leads to a maximum radius of 167 meter for the biggest cluster. The computed spatial clusters were then counted for each user and stored in the measure *num_clusters.* Furthermore, only clusters containing at least 2 points were considered.

4.1.4 Number of Moves, Stops and SSE Points at POI

The computed stops, the SSE points and the move segments are used to compute several measures. The number of moves (*num_moves*) describes the number of moves between SSE points. Therefore, a move segment can be between a start of a trajectory and a stop, between a stop and another stop, or between a stop and an end point of a trajectory. As described in **section 3.6.2**, a move segment is not between an end point and a start point.

Based on that, the average step length, the sum of all covered distances (*mean_step_length*) as well as its standard deviation were calculated (*sd_step_length*). Furthermore, the mean speed in the move segments (*mean_move_speed*) and again its standard deviation (sd_move_speed) were computed.

Since the absolute number of SSE points strongly correlated with the number of moves, only the number of stops itself was considered for further purposes. Besides that, the duration of the stops was taken into consideration. Both the absolute amount of stopping time (*tot_stop_dur*) as well as the mean stopping duration (*mean_stop_dur*) and its standard deviation (*sd_stop_dur*) were calculated.

The SSE points were further used to compute the absolute amount of points that are close to an OSM POI (Table 3.4 in **section 3.4.2**). To test the type of the nearest POI for each SSE point, we measured the distance between each SSE point towards each OSM POI and stored both the distance as well as the type of the nearest POI.

4.1.5 Scenic Roads

Based on information from Tourism Australia (2017), six scenic roads were elaborated to further check for touristic movement. These routes (Table 4.2) are among the routes recommended for self-driving tourists, but have the distinction of not being part of one of the main routes of Australia. For all users, we checked how many days they spent on such routes. The same was done for the two cities of Melbourne and Sydney. The result measures are called *days_scen* (scenic routes), *days_melb* (Melbourne), *days_syd* (Sydney). These measures are only computed for the qualitative interpretation of the clustering results in **section 4.5** and are not used in the machine learning (clustering) part of the methodology.

Table 4.2: The chosen scenic routes as recommended by Tourism Australia (2017)

Name	Start Destination	End Destination	State
Great Ocean Road	Warrnambool	Torquay	Victoria
Great Alpine Road	Wangaratta	Bairnsdale	Victoria
Great Eastern Drive	Hobart	Bay of Fires	Tasmania
Nature's Way	Darwin	Kakadu NP/Litchfield NP	Northern Territory
Gibb River Road	Derby	Wyndham	Western Australia
Upper Gold Coast	Noosa	Bundaberg	Queensland
NSW Coastal Drive	Wollongong	Eden	New South Wales
Uluru-Kata Tjuta National Park			Northern Territory

4.2 Principal Component Analysis

In this study, a set of *36* variables (see Table 4.1) defines each user. From these 36 variables, only the shaded ones (1 to 33) were chosen for the Principal Component Analysis (PCA). Furthermore, only the spatio-temporal footprints of users using the app for at least 5 days were taken into the PCA.

Each user is characterized by the measurements of a 33-dimensional space, built up by the different variables corresponding to its spatio-temporal behavior. It is, however, possible that not every variable is equally important, meaning that it does not contribute a significant amount of additional information. The goal of the PCA is now to find a reduced amount of dimensions that explain as much variability in the data as possible (James et al. 2013).

4.2.1 Variable Transformation

Before applying PCA on a set of variables, it can be advisable to transform and scale the variables. The goal of scaling is to transform the variables in way that they are more comparable to each other. By apply a scaling method, the individual variables will be transformed to an equal scale and therefore have the same variance. Accordingly, every variable gets the same opportunity to be modelled (Bro & Smilde 2003; Bro & Smilde 2014). In our case, we scaled the variables by subtracting the mean of the variable from the original values and divide it by the standard deviation of the variable:

$$\frac{x - mean(x)}{standard\ deviation\ (x)}.$$

Besides scaling, it can sometimes be advisable to transform the variables to correct for heteroscedasticity. A (dependent) variable shows heteroscedasticity when the variability is unequal across the range of values of another, independent second variable that is predicting the first variable. Accordingly, resulting scatterplots will show cone-like shapes. By correcting a variable for heteroscedasticity, we nonlinearly convert the variables in order to make skewed distributions more symmetric (Kvalheim et al. 1994; van den Berg et al. 2006).

To test which transformation is best for the individual variables, a sensitivity analysis was carried out. Each variable was transformed with eight different methods, summarized in Table 4.3 with its corresponding R code formula.

Table 4.3: Applied transformations and their corresponding R formulas

Name of Transformation Method	*Formula*
Log	*log (x+1)*
Log-log	*log (log (x+1) +1)*
Exponential	*exp(x) - 1*
Square-root	*sqrt(x)*
Cube-root	$x^{1/3}$
Squared	x^2
Arcsine	*asin(x)*
Box-Cox	*BoxCox(x, lambda)* whereas *lambda = BoxCox.lambda(x)*

For each of the transformations, the variable was scaled and the skewness value of the distribution was calculated. The transformation which then generated the smallest skewness value was chosen to apply on the variable. An overview of the chosen transformation and its resulting skewness value can be found in Table 4.4. No transformation and scaling was carried out on the variable *num_dstnct_wdays*, since it consists of values between 0 and 7 and can therefore be seen as categorical values. An example of such a transformation is shown in Figure 4.3, depicting the histogram of the variable *tot_dist* before and after a log-transformation and scaling.

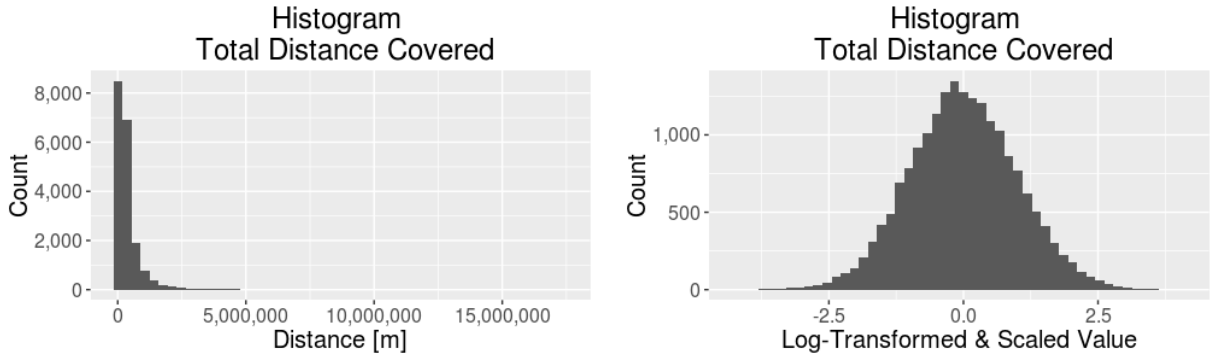

Figure 4.3: Histogram of the variable tot_dist before and after log-transformation and scaling

Table 4.4: Calculated values per user plus their respective transformation and skewness

Variable name	Transformation	Skewness	Variable name	Transformation	Skewness
num_days	Box-Cox	-0.222958	sd_d_cent_dist	Log	0.295666
num_cons_days	Box-Cox	-0.211128	mean_dist_overall_cent	Log	0.341430
num_dstnct_wdays	none		sd_dist_overall_cent	Log	0.244463
period	Box-Cox	-0.204735	num_move	Log-log	-0.019090
tot_dist	Log	0.038953	mean_step_length	Box-Cox	0.304437
tot_time	Log-log	0.116752	sd_step_length	Cube root	0.524025
area	Log	0.100981	mean_move_speed	Box-Cox	-0.170996
circum_hull	Log-log	-0.214989	sd_move_speed	Box-Cox	0.220747
max_dist	Log-log	-0.255113	num_stops	Log-log	0.382013
complex	Log	-0.049242	num_stops_osm	Box-Cox	0.018134
mean_d_dist	Log	0.001933	sse_osm	Log	0.069613
sd_d_dist	Log	-0.154309	tot_stop_dur	Log	-1.801491
mean_d_area	Log	-0.177279	mean_stop_dur	Cube root	1.131058
sd_d_area	Log	-0.139491	sd_stop_dur	Log	-0.282285
mean_d_overlp_pc	Square root	-0.106194	sd_time	Box-Cox	0.129586
sd_d_overlp_pc	Expononential	0.063043	num_clusters	Log-log	-0.042493
mean_d_cent_dist	Log	0.177174			

4.2.2 Interpretation of PCA Outputs

Based on the now transformed and scaled variables, a PCA was carried out. As it can be seen in Figure 4.4 and Table 4.5, the first two components describe a relative big amount of the variance of the original variables. At the third principal component, the proportion of variance drops. At the sixth principle component, the steepness of the curve in Figure 4.4 again drops and is continuing with an even descent.

To visualize the score of each user as well as the loading of each variable on the first principal components, a biplot was generated (Figure 4.5). It shows that a lot of users are centered around the center of the plot [0,0]. Certain users however can be seen in the right bottom of the biplot, therefore having bigger loadings on the first principal component and smaller ones on the second principal component. What can further be seen is that two variables (*mean_move_speed, mean_step_length*) show into the opposite direction of most of the other ones. Already in this plot, that only shows the first two principal components, a certain kind of clustering can be seen by having one large group around the center and one smaller in the left bottom corner.

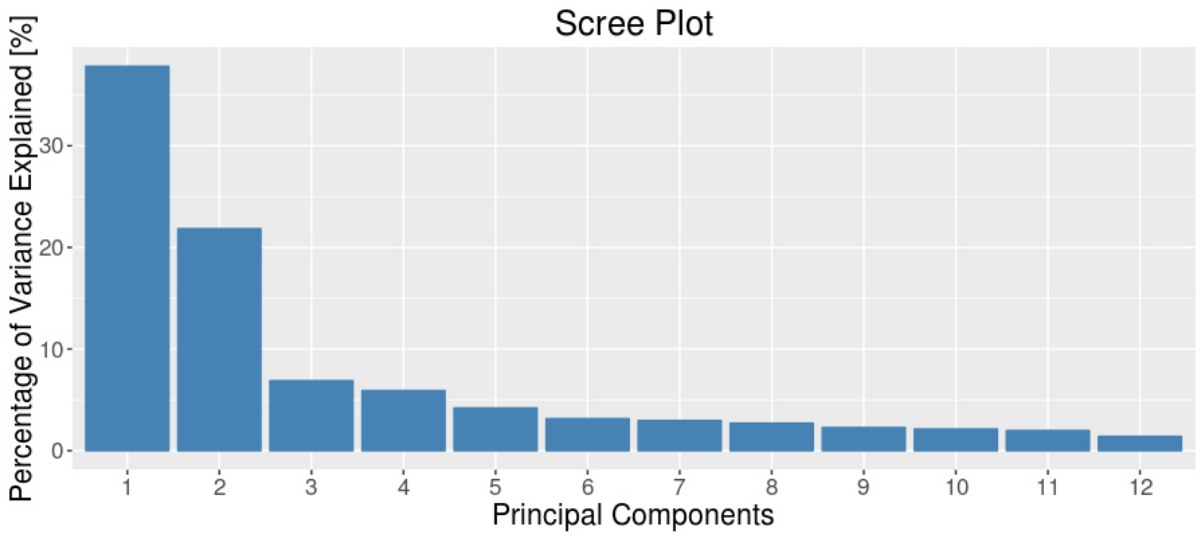

Figure 4.4: Percentage of variance explained per principal component.

Table 4.5: Importance of the individual components of the PCA

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10
Standard deviation	3.462	2.614	1.464	1.346	1.154	1.003	0.971	0.872	0.832	0.824
Proportion of Variance explained [%]	38.66	22.03	6.91	5.84	4.30	3.25	3.04	2.46	2.34	2.12
Cumulative Proportion of Variance explained [%]	38.66	60.69	67.6	73.44	77.74	80.99	84.03	86.49	88.72	89.82

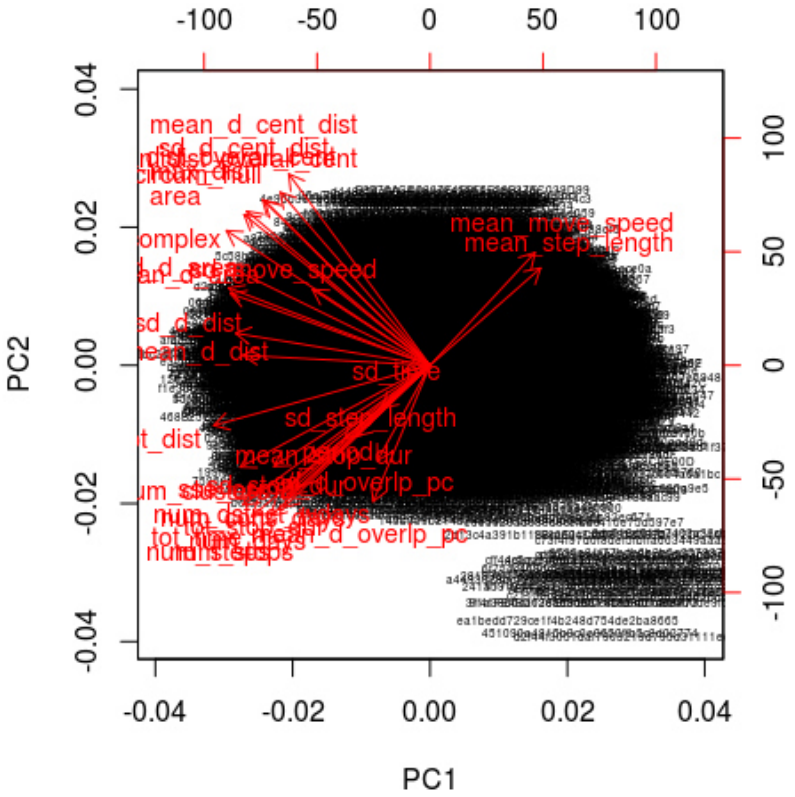

Figure 4.5: Biplot of the first two components of the PCA

4.2.3 Description of the found Principal Components

To get a deeper understanding of the found principle components, the following sections will give an overview of the first four principal components as well as a short interpretation.

First Principal Component

The first principal component PC1 (Figure 4.6) is dominated by a variety of input variables whereas the variables *tot_distance, complex* and *area* contribute to PC1 with the highest percentage values. If we compare the PC1 values with the values of the original variables, we can state that the higher the PC1 value…

…the smaller the total distance covered,

…the smaller the complexity (overall area of concave hull divided by circumference),

…the smaller the overall area covered,

…the smaller the average and the standard deviation of the daily area covered and

…the smaller the standard deviation of the daily distance covered.

Based on these observations, several interpretations can be made. Users with a rather high PC1 value tend to cover big distances and a wide range of different trips, thus covering large areas. The first principal component can therefore be interpreted as some sort of description of the overall mobility of the individual users.

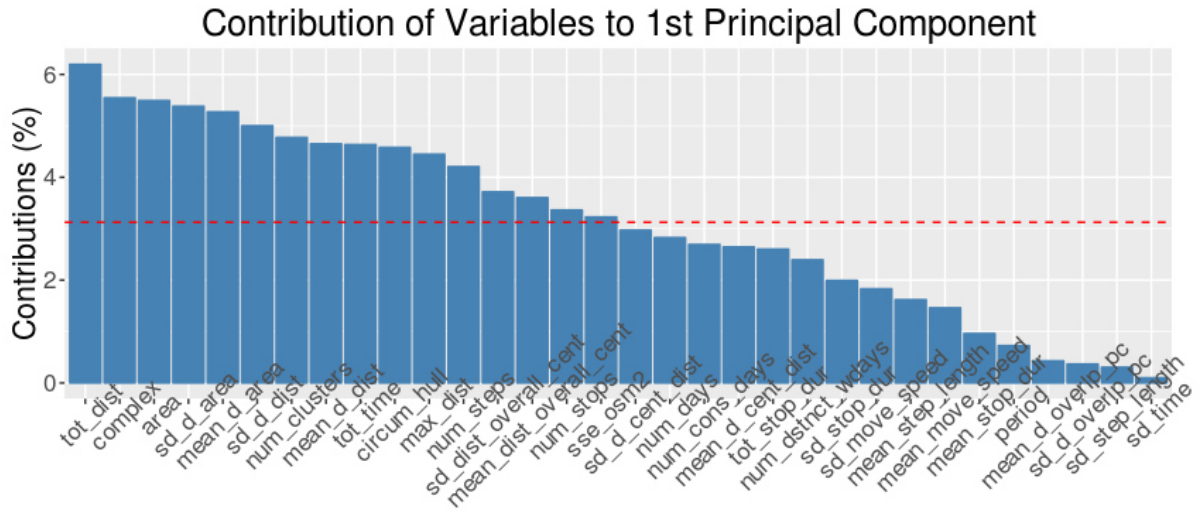

Figure 4.6: Contributions of the individual variables to the first principal component

Second Principal Component

The second principal component (Figure 4.7) is highly dominated by the contributions of the distances of two consecutive centroids (mean and standard deviation) as well as the distance of the daily centroids to the overall centroid (mean and standard deviation). Other variables with high contributions are the number of stops and the number of moves. If we compare the PC2 values with the original variable values, we state that the higher the PC2 value…

…the higher the average and the standard deviation of the distance between two consecutive daily centroids,

…the higher the average and the standard deviation of the distance between the overall centroid and the daily centroids,

…the higher the maximum distance in the overall concave hull,

…the smaller the number of stops and

…the smaller the number of moves.

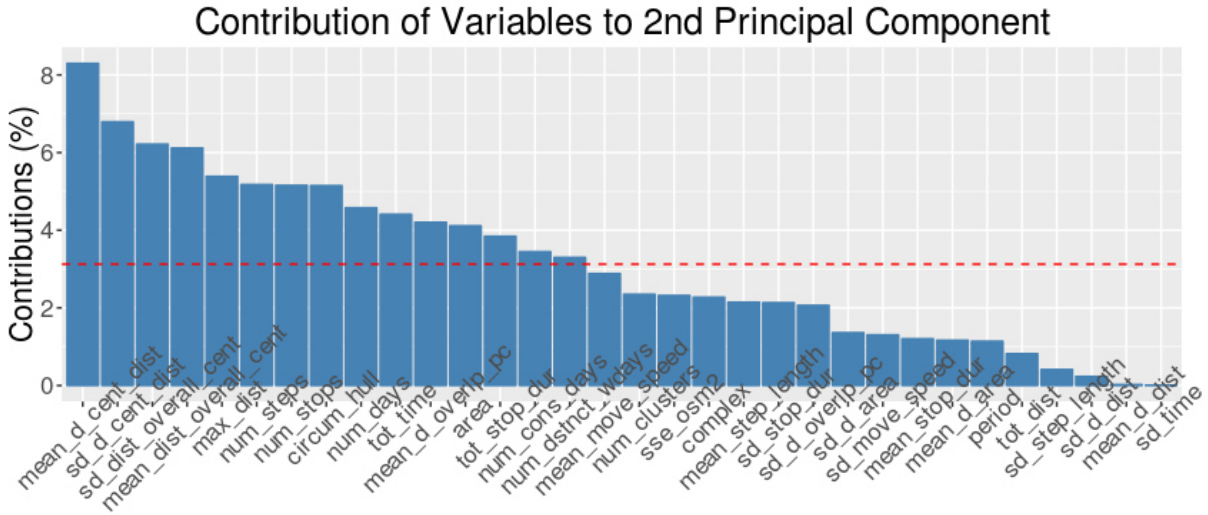

Figure 4.7: Contributions of the individual variables to the second principal component

Based on these observations, the second principal component can be defined as a description of the spatial inconsistency of the individual users. Accordingly, the smaller the spatial inconsistency (PC2), the higher is the probability that the user has spent its time in the same environment, the same city. The higher the spatial inconsistency (PC2), the higher is the probability that the user has moved around a lot.

Third Principal Component

The third principal component is dominated by the contributions of the average daily distance covered (*mean_d_dist*) and its standard deviation (*sd_d_dist*) as well as the mean overlap of the concave hull of two consecutive days (*mean_d_overlp_pc*) and its standard deviation (*sd_d_overlp_pc*). If we compare the PC3 values with the original variable values, it can be stated that the higher the PC3 value…

…the smaller the average and the standard deviation of the daily distance covered and

…the higher the average and the standard deviation of the overlap of the consecutive concave hulls.

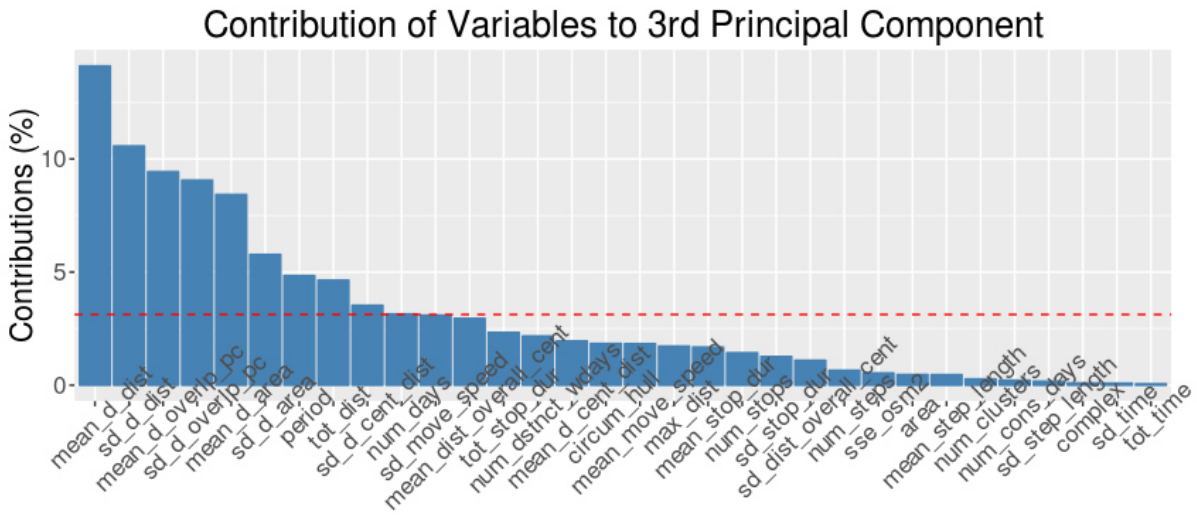

Figure 4.8: Contributions of the individual variables to the third principal component

Based on these observations, the third principal component can be seen as the opposite of PC2 and describes the spatial consistency of a user. Accordingly, the higher the spatial consistency (PC3), the higher is the probability that the user has spent its time in the same environment.

Fourth Principal Component

The forth principal component is highly dominated by the contribution of the average stop duration (*mean_stop_dur*). As it can be seen in Figure 4.9, smaller, but still important contributions are made by the standard deviation of the stop duration as well as the overall cumulated stop duration. If we compare the PC4 values with the original variable values, it can be stated that the higher the PC4 value…

…the higher the average and the standard deviation of the stop duration and

…the higher cumulative stop duration.

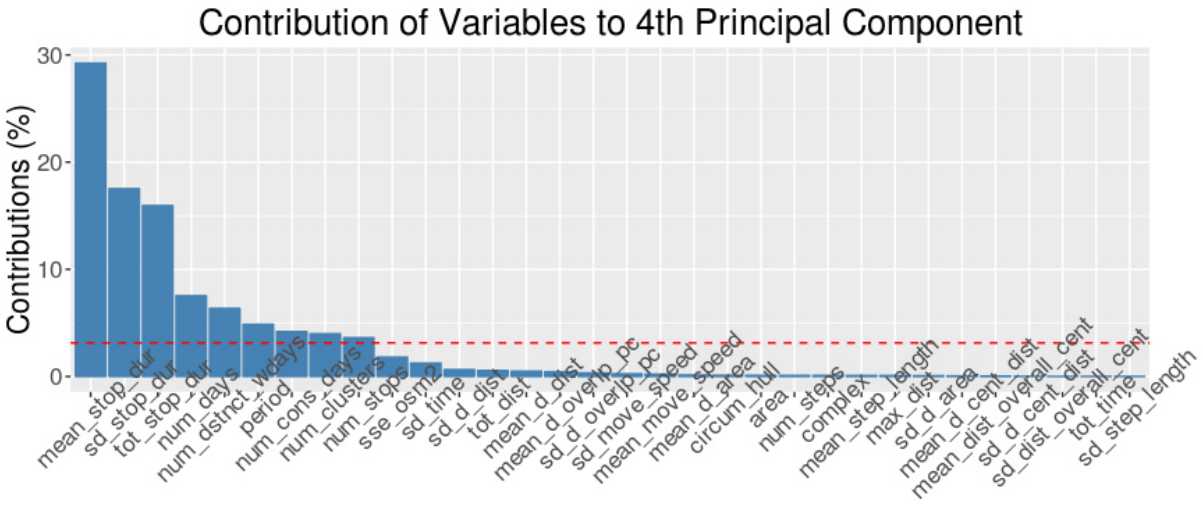

Figure 4.9: Contributions of the individual variables to the fourth principal component

Based on these observations, the forth principal component can be described as the stopping behavior of a certain user. Accordingly, the higher the stopping behavior (PC4), the higher is the probability that the user shows a lot of stops in his trajectories.

4.2.4 Choosing the Number of Principal Components

As reviewed in the background section, there are several methods to choose the number of principal components. Accordingly, the scree plot as well as the eigenvalue method and the broken stick method were applied to come up with reasonable numbers of principle components.

In Figure 4.10, all three methods combined are presented. It shows that the elbow of the scree plot is at the third principal component. However, this is a very small number of components that only explains about 67.601% of the variation (see Figure 4.11). Regarding the eigenvalue, Figure 4.10 shows that the first six principal components have an eigenvalue bigger than 1 (Kaiser-Guttman's criterion), explaining 80.99% of the variation. The last applied method, the broken stick method, also indicates that three principal components seem to be reasonable.

As Bro & Smilde (2014) suggest, it may be reasonable to take more than just two or three components in case they only describe 50% of the variance. On the other hand, it can lead to an overfitting when choosing all components that describe 90% of the variance, but there is a lot of noise in the data. In the here presented case (see Figure 4.11), the first two components describe 59.26 % of the variance whereas about 90% are described by the first ten components.

Based on the recommendations found in Bro & Smilde (2014), three different numbers of principal components were chosen and taken into the clustering step: three, four and six principal components. Three principle components where chosen based on the knee in the scree plot and the interception of the broken stick distribution with the scree plot. Six principle components have an eigenvalue bigger than 1 and therefore 6 PC's were additionally chosen for further analysis. Finally, four principle components were chosen, since the fourth principle component has high contributions from the individual stopping

behavior, which may lead to a separation of users with a rather touristic behavior that stop more on their routes.

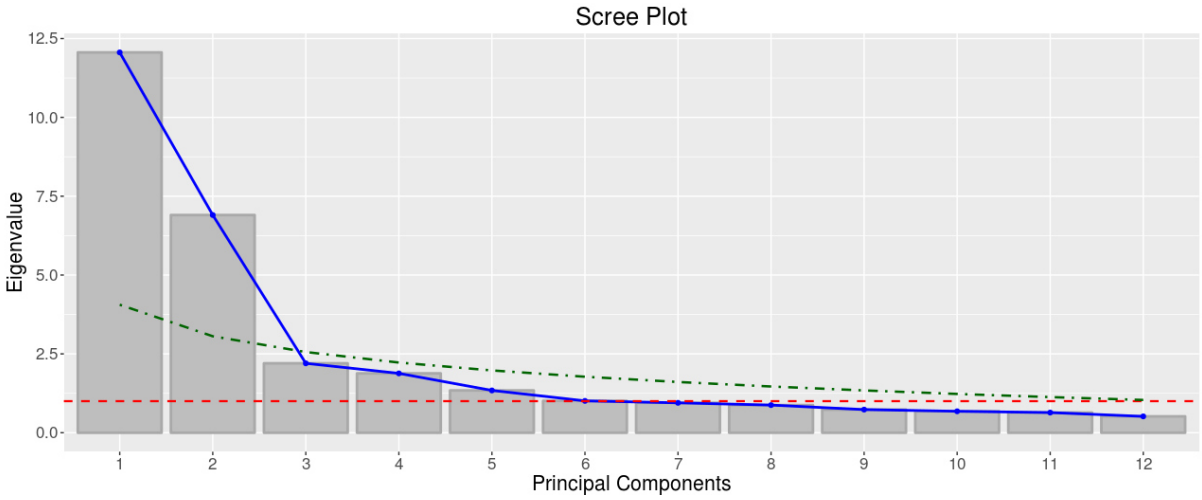

Figure 4.10: Scree plot (blue) of the first twelve components, with the line where the eigenvalue = 1 (dotted red) and the broken stick distribution (dotted green)

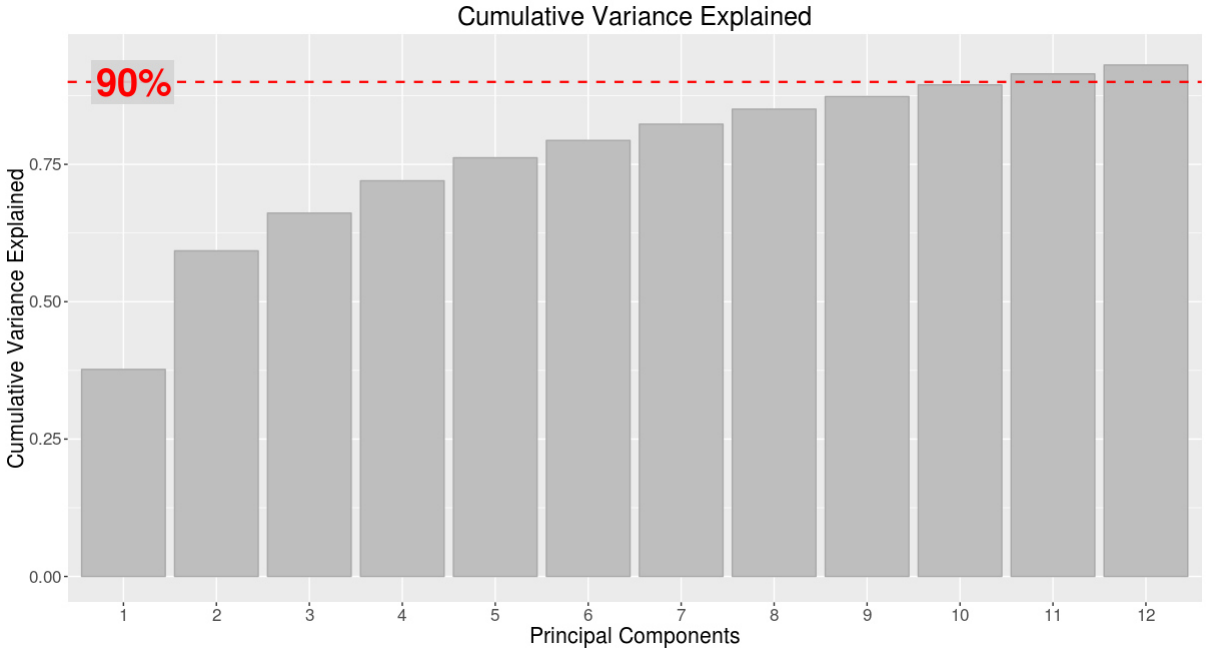

Figure 4.11: Cumulative variance explained per principal component, 90%-line in red

4.3 User Categorization through Clustering

The division of users into several classes based on their principal components can be made in many ways as described in **sections 2.3** and **2.4**. In this case, no a priori knowledge of the class labels is given. Accordingly, an unsupervised learning technique such as clustering must be applied. Clustering is a process of categorizing the individual objects into groups; unlike the classification approach however, it creates the class labels only based on the data given (Tan et al. 2006). Within the individual groups created by the clustering algorithm, the objects should have a high similarity. Among the groups, however, there should be a high dissimilarity (Han et al. 2012).

Based on literature review, we explored the performance of four different clustering algorithms which were each presented in the literature review in section 2.4: K-Means, CLARA, AGNES, and DIANA.

4.3.1 Cluster Validation

In this section, three numbers of principal components (three, four and six) will be tested with three clustering validation techniques (silhouette width, gap statistic, stability measures) for four different clustering methods (K-Means, CLARA, AGNES and DIANA). As suggested by various studies (Han et al. 2012; James et al. 2013), the chosen principle components have been scaled before applying a cluster method.

Silhouette Width

The results of the silhouette width for the four different clustering algorithms and for three different numbers of components chosen are shown the top plots of Figure 4.12. A comparison of the computed silhouette widths for three principle components (top left) shows that AGNES produces the best values, i.e. the best clustering, for up to five clusters. The silhouette width for AGNES is, however, declining with the number of clusters chosen. The second-best clustering algorithm with three principle components is K-Means with equal silhouette widths for all number of clusters, however, highest with three clusters. As in the other approaches with different numbers of principal components as well, DIANA generally generates the lowest silhouette widths.

For four principle components, again AGNES has the highest silhouette widths for up to four clusters. From five clusters on, the then best approach is always given by using K-Means, whereas the highest silhouette width is given for five clusters. In the approaches with six principle components, again AGNES produces the highest silhouette widths for all different cluster numbers.

Gap Statistic

The result of the gap statistic for different numbers of components with four different clustering methods can be seen in the bottom plots of Figure 4.12. For all numbers of clusters, DIANA and AGNES produce a far worse gap statistic value than K-Means and CLARA. Especially the result of AGNES stands in a strong contrast to the results of the silhouette width, in which AGNES produced by far the best results. Regarding the actual gap statistic value, the bottom left plot in Figure 4.12 shows that for three principle components, CLARA and K-Means produce the highest value with 1 cluster, followed by 3 and 4 clusters.

For four principal components (top right), the highest gap statistic value is given for CLARA and 5 clusters. In the approach with 6 principle components, again CLARA and K-Means with only 1 cluster produce the best gap statistic value, followed by CLARA with 8 clusters.

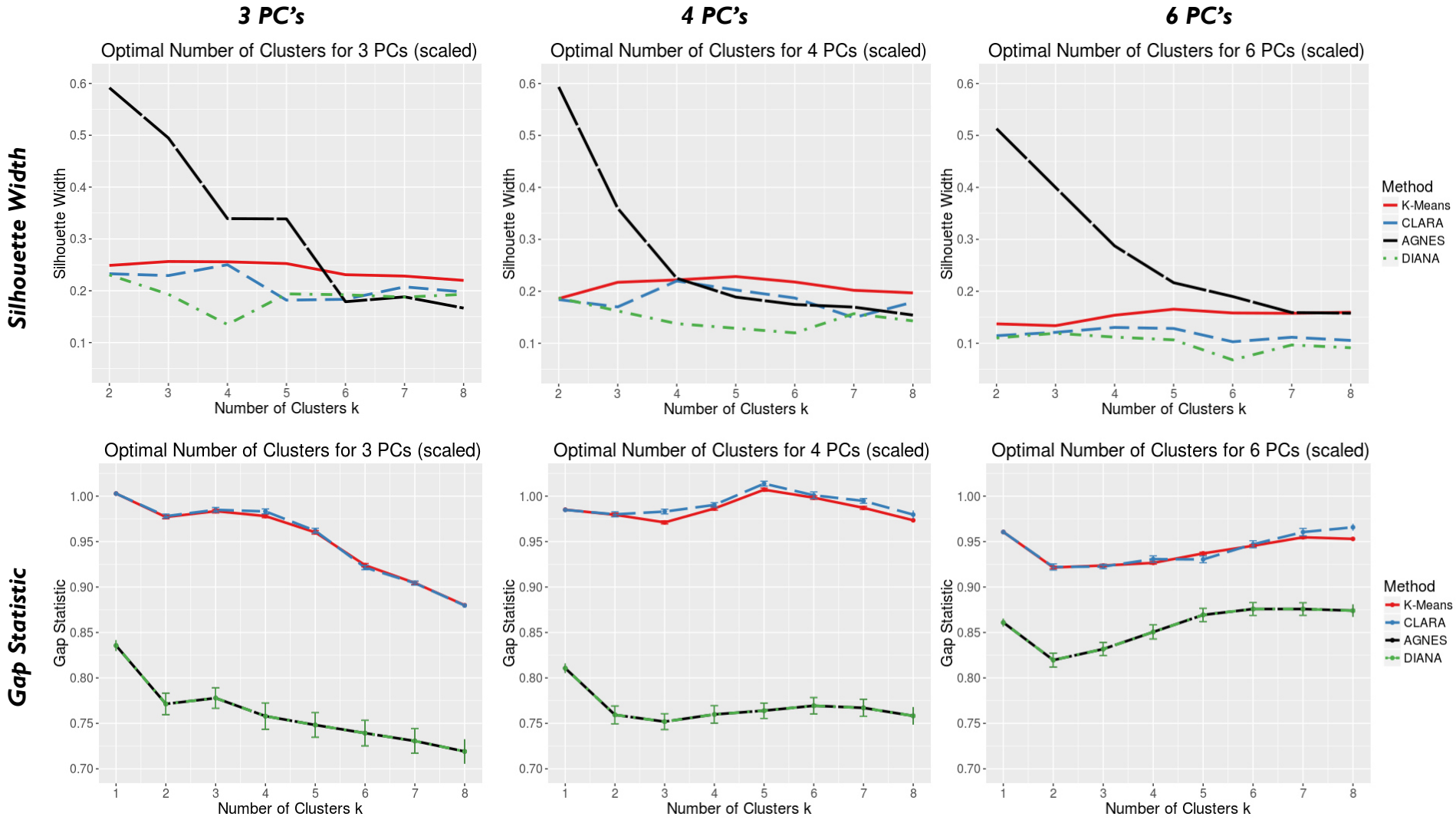

Figure 4.12: Top: Silhouette width for four different clustering methods based on the first three (left), first four (middle) and top six (right) principal components, scaled. Bottom: Gap statistic for four different clustering methods based on the first three (left), first four (middle) and top six (right) principal components, scaled.

Stability Measures

In Figure 4.13, the three stability measures (APN, AD & ADM; description in **section 2.4.1**) are plotted for the four different clustering methods and different numbers of clusters based on three (left), four (middle) and six (right) principle components.

The plots for the APN (top) and ADM (bottom) stability measures show that the better the clustering results, the lower the number of clusters and number of principle components we consider. The plots further produce very comparable results for all numbers of principle components, number of clusters and clustering methods respectively. For three principle components, both the best APN and ADM values (the lower the better) are given for AGNES with either two, three or four clusters. On the contrast, the worst APN and ADM values are presented by CLARA with six clusters. For four principle components, the best and lowest APN and ADM values are again for AGNES, but with only two clusters. The worst values on the contrary are again presented for a clustering with CLARA. For six principle components, the same can be seen as for the other two principle components. Again, AGNES produces the best results and CLARA the worst.

The AD plots (middle) show a different result. In contrast to the other two stability measures, we get the best clustering when considering more numbers of clusters, but a smaller amount of principle components. For all principle components and clusters, AGNES produces the worst, i.e. the highest values, whereas K-Means produces the best clusterings.

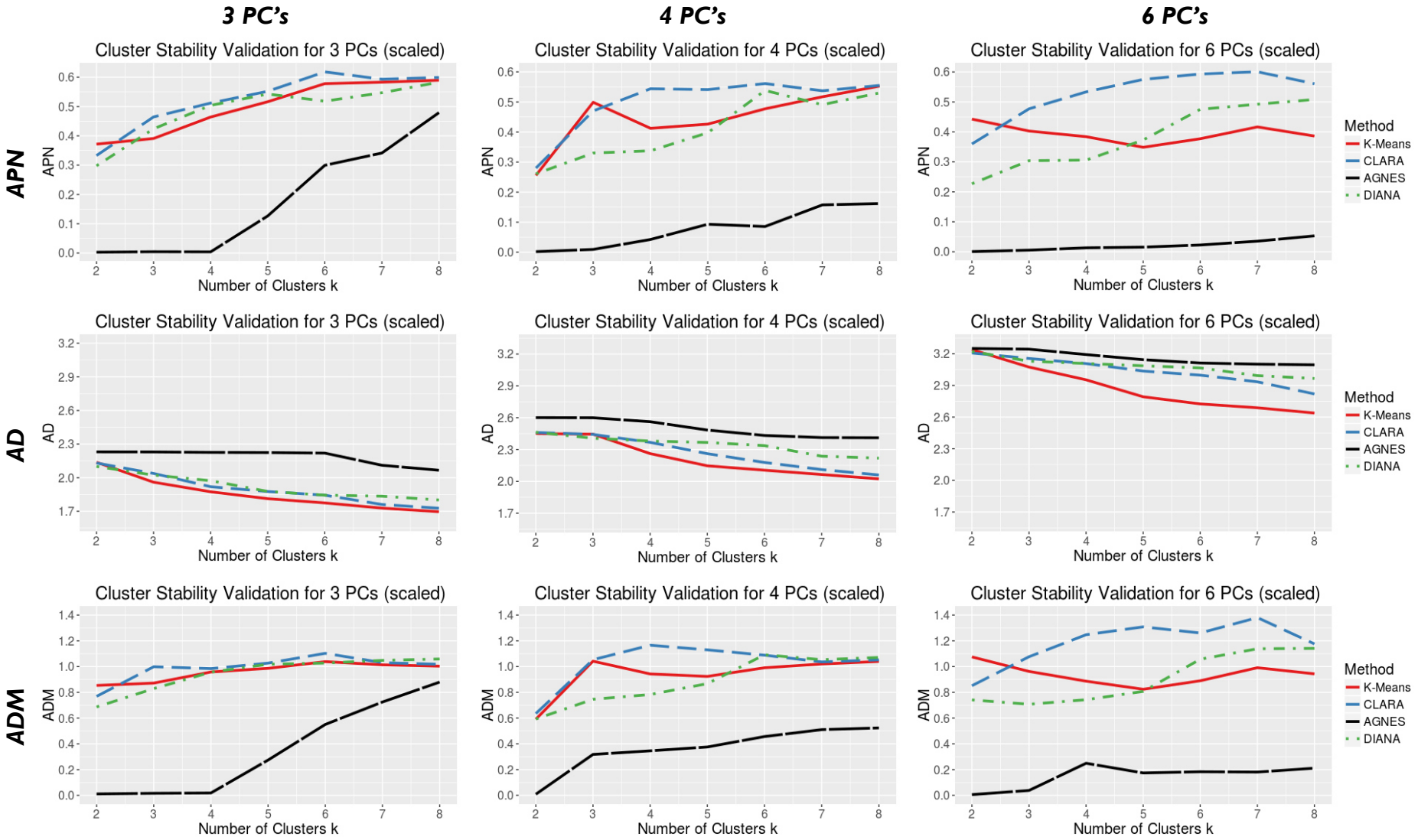

Figure 4.13: Stability measures per clustering method for three PCs (left), four PCs (middle) and six PCs (right); all PCs scaled.

4.3.2 Rank Aggregation

When looking at the biplot from the first two principal components (Figure 4.5 in section 4.2.2), certainly more than only one cluster can be identified. Although there is a big cluster of users in the center of the plot, also a small cluster in the right bottom corner (high values of PC1 and small values of PC2) can be seen.

Based on the results of the silhouette, gap statistic and stability validation plots, however, it is difficult to come up with a solution for both the questions regarding a good amount of PC's and the ideal clustering algorithm. We therefore use rank aggregation as proposed by Pihur et al. (2009) to come up with the best approach.

Rank aggregation aggregates a combination of ranked lists to generate an overall ranking. Accordingly, the goal of rank aggregation is to find a so called "super"-list which shows the highest consistence with all individual lists simultaneously. To measure the distances between the ranked lists, several approaches can be used, such as Spearman footrule distance and Kendall's tau distance (Pihur et al. 2009).

The Spearman footrule distance between two lists can be described as the sum of the absolute distances between the ranks of all the unique elements from the two ordered lists. In case of comparing clusterings, not only ranks are given, but also respective measures, such as the gap statistic for each clustering. Accordingly, Pihur et al. (2009) came up with the weighted Spearman footrule distance (WSF) which uses the values such as the gap statistic as additional weightings.

The other proposed approach is to measure the distance between ranks is the Kendall's tau distance (Pihur et al. 2009). Like WSF it uses pairs of elements from two lists and compares their rank. If an element does not have the same rank in the two lists, a penalty is imposed. In the proposed algorithm by Pihur et al. (2009), the weighted Kendall's tau distance (WKT) uses a penalty value that is dependent on the absolute difference in an element's scores from two different lists. To compute the aggregated list, Pihur et al. (2009) use the Cross-Entropy Monte Carlo Algorithm (CE). Besides the ranking and the scored, CE can also take an importance rating for each list as an input variable.

In the given case, three numbers of principal components, four clustering algorithms and five numbers of clusters (2 to 6 clusters) generate 60 values for five different measures respectively (silhouette, gap, AD, ADM, and APN). Accordingly, we are presented with five ranked lists and their respective scores. Since the results of the APN and the ADM statistic (see top and bottom plots in Figure 4.13) show the same pattern, we weigh these two variables equally. Moreover, we weigh all three stability measures combined as equal as the gap statistic and the silhouette width due to the diversity, respectively similarity of the measures. Accordingly, following importance ratings were applied: silhouette (1), gap (1), APN (0.25), AD (0.5) and ADM (0.25).

The CE was applied twice, once with WSF and once with WKT. The weighted ranking with WSF as distance measure showed that the approach with three principal components and K-Means with three clusters produces the highest score (see Figure 4.14), followed by three principal components and AGNES with three clusters (Table 4.6).

The grey lines in Figure 4.14 show the ranks of each approach for the individual clustering validation measures, whereas the black line shows the mean rank of each approach. The red line shows the overall ranking of each approach based on the CE algorithm. Accordingly, the best ranked approach can be seen on the far left and the worst approach on the far right.

Table 4.6: The top 5 approaches based on the Cross-Entropy Monte Carlo Algorithm with the Spearman footrule distance (left) and the Kendall's tau distance (right)

Top 5 Spearman footrule distance			Top 5 Kendall's tau distance		
Number of PCs	Clustering Method	Number of Clusters	Number of PCs	Clustering Method	Number of Clusters
3	K-Means	3	3	K-Means	3
3	AGNES	2	4	K-Means	5
4	K-Means	5	4	K-Means	4
3	K-Means	4	3	K-Means	4
3	CLARA	2	3	CLARA	2

Using rank aggregation, we conclude that a clustering into three clusters with K-Means based on three principle components (3PC-X3KM) is the most appropriate clustering method. When considering the aggregated ranking with the WSF, the second rank goes to AGNES with 2 clusters based on three principal components whereas Kendall's tau distance leads to five clusters with K-Means based on four principal components (4PC-X5KM). Since this approach takes the third rank in the WSF approach, 4PC-X5KM can be considered as the second-best approach overall.

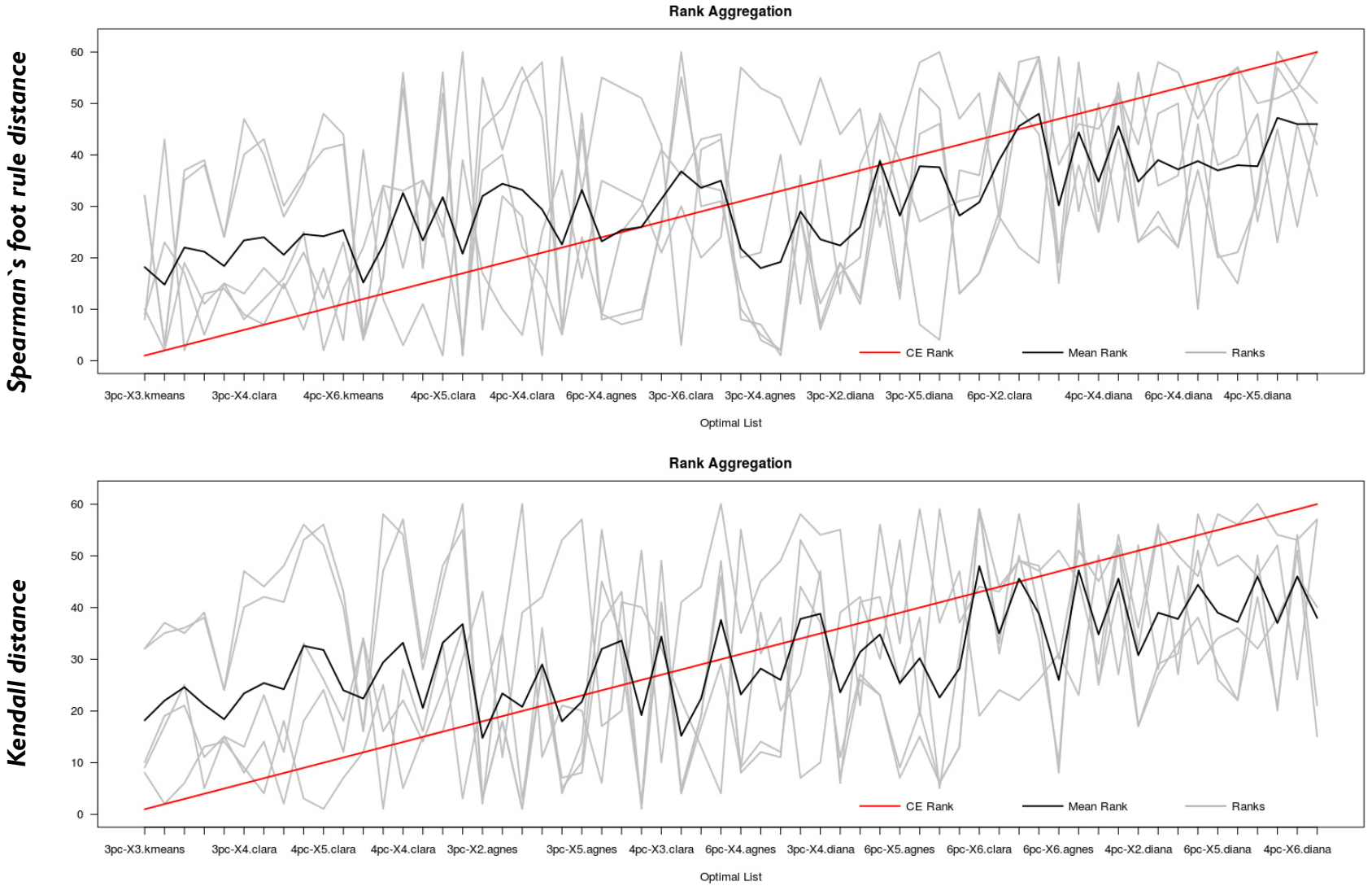

Figure 4.14: Rank aggregation for Spearman's foot rule distance (top), Kendall distance (bottom), 3 different numbers of PCs, 4 clustering methods and 5 numbers of clusters (two to six)

4.4 Selection of Clustering Approach

The goal of our clustering approach is both to clearly distinguish between different kinds of users as well as to get groups that can be put into context. According to James et al. (2013), a good clustering has both good statistical properties as well as useful and interpretable solutions. The goal of this approach is to select certain clusters and put them into context, without having additional ground truth to refer to. Due to that, we take the original user measures for each cluster and use them to interpret the found clusters. They therefore serve as ground truth used to form the different user types.

The rank aggregation of the various clustering approaches in **section 4.3.2** has led to two rankings, headed by the 3PC-X3KM and the 4PC-X5KM approach. Due to the closeness of these two first approaches in the two rankings, we decided to do a first small interpretation of the two best-ranked clustering approaches to find the best-suiting among them. Accordingly, we define best-suiting not only in terms of ranking, but furthermore in the interpretability of the various found clusters, as suggested by James et al. (2013).

To compare the two approaches, we use three boxplots and compare the distributions of these three original values across the cluster groups (Figure 4.15). By having this first visual assessment of the diversity in the individual clusters, we select the best-suiting approach.

Boxplot Interpretations

The 3PC-X3KM-approach (left boxplots in Figure 4.15) presents us with three different clusters, whereas the 4PC-X5KM is presenting five clusters (boxplots on the right). As the boxplots for the first chosen variable, average daily distance, show there is not much variation between the different 3PC-X3KM-clusters, whereas for 4PC-X5KM, much more variety between the different clusters can be seen.

In the second presented case for the average daily overlap, a similar image is shown, however, not as extreme as in the first example. The same can be stated for the third example for the number of stops. In both cases, the clusters in the 4PC-X5KM-case can be better separated from each other than in the 3PC-X3KM-case.

The main goal of the clustering approach is to get meaningful groups that are interpretable. Based on the three chosen variables, we believe that the 4PC-X5KM-clusters are better interpretable than the 3PC-X3KM-clusters, due to the bigger variance among the clusters. We therefore decided to proceed with the 4PC-X5KM-clusters and neglect the better-ranked 3PC-X3KM-approach.

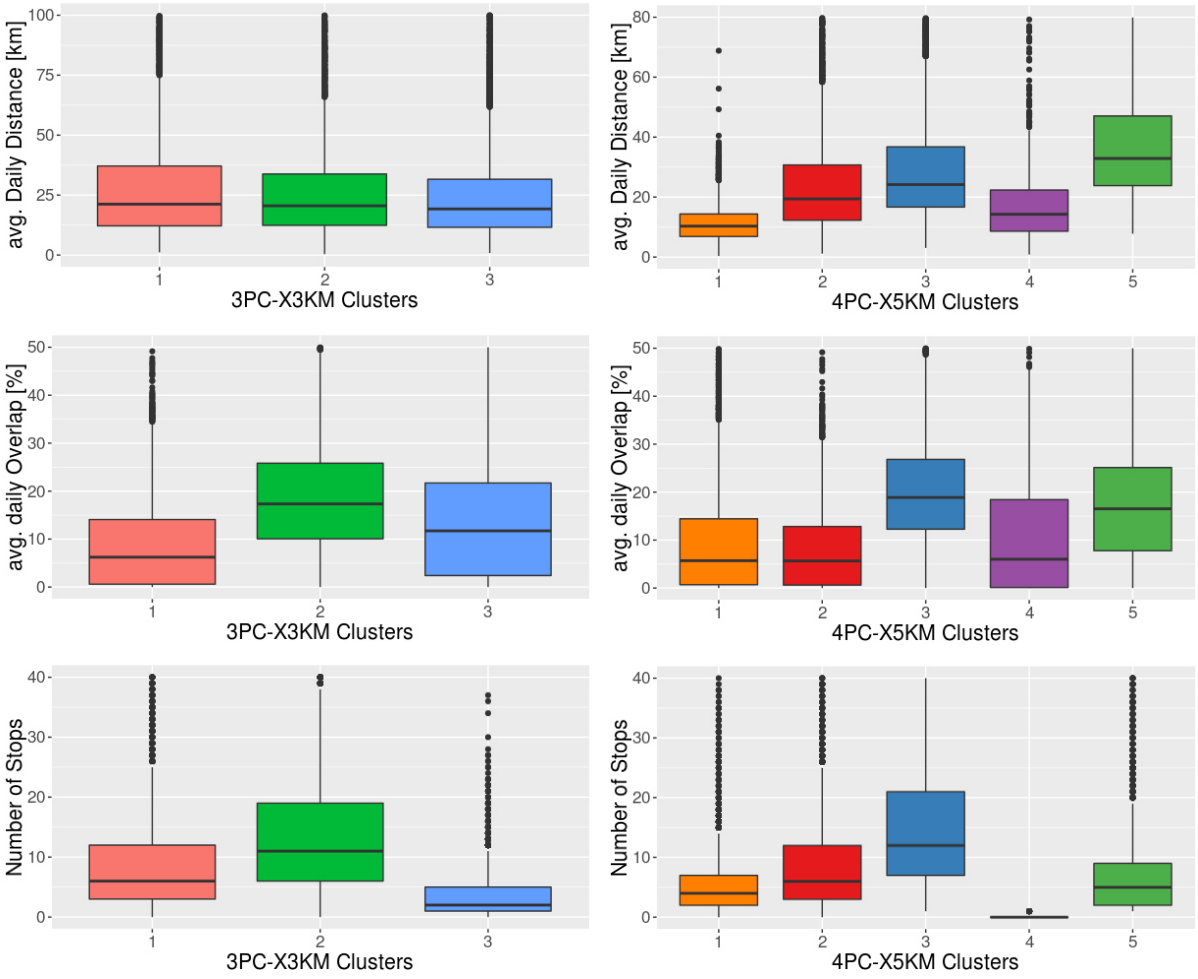

Figure 4.15: Boxplots of the average daily distance (top), average daily overlap (middle) and the number of stops for the two clustering approaches (3PC-X3KM: left, 4PC-X5KM: right) and their clusters

4.5 From Clusters to User Types

In the following sections, we are trying to label the different clusters of the chosen 4PC-X5KM-approach. We are, however, aware of the fact that putting labels on individual clusters is rather difficult due to the lack of ground truth. Accordingly, the labels used for the user types are only an assumption based on the characteristics of the clusters.

We therefore proceed as follows in the following sections. Firstly, the individual clusters will be described based on their manifestations in the original values, using boxplots found in Figure 4.16 and Figure 4.17. Secondly, the findings of the descriptions will be used to interpret the clusters. Based on the interpretation we then assign a user type to each individual cluster. We are starting with the description and interpretation of the best interpretable clusters and are therefore not following the actual cluster numbers.

4.5.1 Cluster 2/User Type T: The "Tourist" / The "Overland Delivery Driver"

Description

Cluster 2 (size: **3'343**) users show a small proportion of days spent in Sydney/Melbourne (median value: 20%), combined with a rather large number of days in total (10d) and a high total area covered by the trajectories (1'600km^2). Conversely, cluster 2 users have a very small compactness value (0.04). While their proportion of days in the two cities is small, also their proportion of days spending on one of the scenic routes is small (15%). Cluster 2 users strongly contrast with the other clusters with both their high average centroid distance (150km), the high average distance between the daily and the overall centroid (80km) and especially the distance between the two most distant points (450km). The overall distance (200km) and the average daily distance (20km) are about average. Rather smaller values are given for the average daily overlap (6%), move speed and step length (480m). The number of stops (12) as well as the number of spatial clusters (30) then are rather higher than most of the other clusters, except cluster 3.

Interpretation

Cluster 2 users show a special pattern. Due to their high total area covered but rather average total distance, they are hard to put a label on. Their rather small average overlap reflects perfectly their spatial instability. Accordingly, there are various possible explanations for their spatio-temporal footprints. Firstly, the user type of cluster 2 may reflect the behavior of someone who travels around the country such as, for example, an overland delivery driver. That type of user visits a lot of different regions which leads to a high area value. We therefore assume that these types of users have a rather repetitive behavior, may be re-visiting the same places again over a certain amount of time. Due to that, the driver may already know some of the stretches and does turn off the device for some time, which leads to a relatively small total distance covered.

The second possibility is that this user type may reflect a certain kind of touristic behavior where the tourist visits different regions of the country which leads to a big area. To move between the regions, however, he uses alternative means of travel such as planes which can lead to only an average amount of total and daily distance covered.

Based on the interpretations, we can give users of cluster 2 two different labels. Firstly, the one of the "overland delivery driver" due to the first possibility described before. Secondly, we can label cluster 2 as touristic user type. We therefore assign cluster 2 users the **user type T**, based on the possible touristic behavior.

4.5.2 Cluster 3/User Type C: The "Commuter"

Description

Cluster 3 users (size: **5'861**) differ from other clusters due to their high activity. Their number of days in total (15, IQR: 13-22), their proportion of days in either Sydney or Melbourne (75%), their total distance (400km), their high average daily overlap (20%) and their high number of spatial clusters (42), for which cluster 3 users have the highest

values. Rather high compared to users of other clusters is the degree of compactness (0.2). On the opposite, the lowest values are given for cluster 3 for the average step length (420m), the average centroid distance (12km), the move speed, the proportion of days on scenic routes (10%). About average are the distances between the two most distant points (60km) and the total area (500km^2).

Interpretation

For cluster 3, we assume that users are spending their days in and/or around a bigger city and use the app on a regular basis, i.e. almost daily. This assumption can be made based on several observations. First, cluster 3 users do not have a big average distance between the daily centroid and the overall centroid. Furthermore, they have relatively high spatial overlap which leaves us with the assumption that cluster 3 users have a rather stable spatial behavior. Second, the average speed is low which may can indicate an environment with more traffic. Third, the step lengths are smaller than the ones of other groups which may be related to the smaller distances between their points of interests in the city. This is further reflected in the relative compactness of the concave hulls of their movements. Due to these observations and interpretations, we see a commuting behavior in cluster 3 users and, accordingly, label users of cluster 3 as commuters. Based on these labels, we assign cluster 3 users the **user type C.**

4.5.3 Cluster 5/User Type E: The "Excursionist"

Description

Cluster 5 users (size: **4'372**) show about an average usage, seen in the number of days (5d). While using the app, they tend to visit stretches of the scenic roads (21%) more frequently than users from the other clusters. The proportion of days they spend in either Sydney or Melbourne are about average (40%). For the average daily distance (70km), cluster 5 users have the highest values of all clusters. Second but highest values are given for the total distance (250km), the total area (630km^2), the average daily overlap (17%), the average centroid distance (20km), the average distance between the daily and the overall centroid (18km) and the average move speed. All the other values are among the average compared to the other clusters. The compactness of the concave hull, however, which is about average as well, is by far not as small as the one of cluster 2, but almost as big as the ones from clusters 1 and 4.

Interpretation

Cluster 5 users show a pattern that can be interpreted as touristic behavior. The high average daily overlap value, however, stands in great contrast to this and gives us important additional information, namely that these users have spent at least some time at the same place.

Cluster 5 users may reflect a behavior we call "part-time" or "weekend"-tourist. January is not only one of the main months for tourists from overseas to visit Australia, it is also the month with probably the highest domestic tourism due to summer holidays (see Figure 3.3). Due to that, we assume that a relative large amount of people do not work for

the whole month, but go on holidays or weekend trips. This would also reflect the relative high spatial stability (high average daily overlap, high compactness).

Several further interpretations can be made. A first possibility is that cluster 5 users spent about half their time at their home location and then do a road trip. A second possibility is that the user type stands for someone who works during the week, but does a lot of weekend trips to various places. This may include trips to, for example, beach houses or shopping malls, that are a bit further away from their residence. Accordingly, we call the cluster 5 users the "excursionist" type and assign cluster 5 users the **user type E.**

4.5.4 Clusters 1 and 4: The "small-scale"-Users

Description

Cluster 1 users (size: **4'451**) and **cluster 4** users (size: **1'079**) do not differ much from each other. Users assigned to either of the two clusters have a rather small number of days of usage and therefore not very informative and sound data. They both have the smallest distance between the two most distant points regarding all cluster as well as the smallest total and average distance. Most of the users' measures are either relatively small or among average when compared to the ones of the other clusters. Differences between the two clusters can be seen when looking at the average move speed and the number of stops. Cluster 4 users both have a higher average speed in the move segments than cluster 1 users, but mostly do not show any stops in their trajectories.

Interpretation

Due to their small-scale usage in both space and time, it is difficult to interpret the two clusters. Accordingly, we label them as "small-scale"-users which reflects their behavior of only using the app over small amounts of time and distances. We further do not assign a user type to these users of these cluster, due to the difficult interpretability of both clusters.

4.5.5 Cluster/User Type Description Summary

At least three distinct types of users can be identified using both the boxplots and their corresponding interpretations. The two clusters that have not assigned a user type show very similar behavior and, moreover, only have very small usage values. Accordingly, we decided to only take the three clusters, i.e. user types that have the best interpretability to take forward into the next steps. This means that we are only investigating the temporal and spatial characteristics of these three user types, therefore neglecting the patterns of clusters 1 and 4.

A further investigation into some of the key figures (Table 4.7) reveals some additional information for the found user types. The commuter user type C has by far the largest number of users, especially within the two cities. Type C further shows the largest amount of recorded sessions, along with the largest number of SSE points, both in total as well as in the two cities. In contrast to this stands the percentage of users per user

type that have visited one of the two cities. Here, user type T shows the biggest value with about 64%.

Table 4.7: Summary of the three established user types. Highlighted are the highest numbers among the three user types, respectively.

	Cluster 2	Cluster 3	Cluster 5
User Type	**User Type T**	**User Type C**	**User Type E**
Interpretation	The "Tourist" The "Overland Delivery Driver"	The "Commuter"	The "Excursionist"
Number of Users in Total	3'343	5'861	4'372
Number of Users in Sydney/Melbourne	2'140	3'321	2'572
Percentage of Users visiting Cities	64.01%	56.66%	58.83%
Number of recorded Sessions	70'941	255'584	67'265
Number of SSE Points	178'820	676'723	171'740
Number of SSE Points in Sydney/Melbourne	58'575	337'379	72'925
Percentage of SSE Points in Cities	32.76%	49.85%	42.46%
Number of SSE Points per Session	2.520	2.647	2.553
Color			

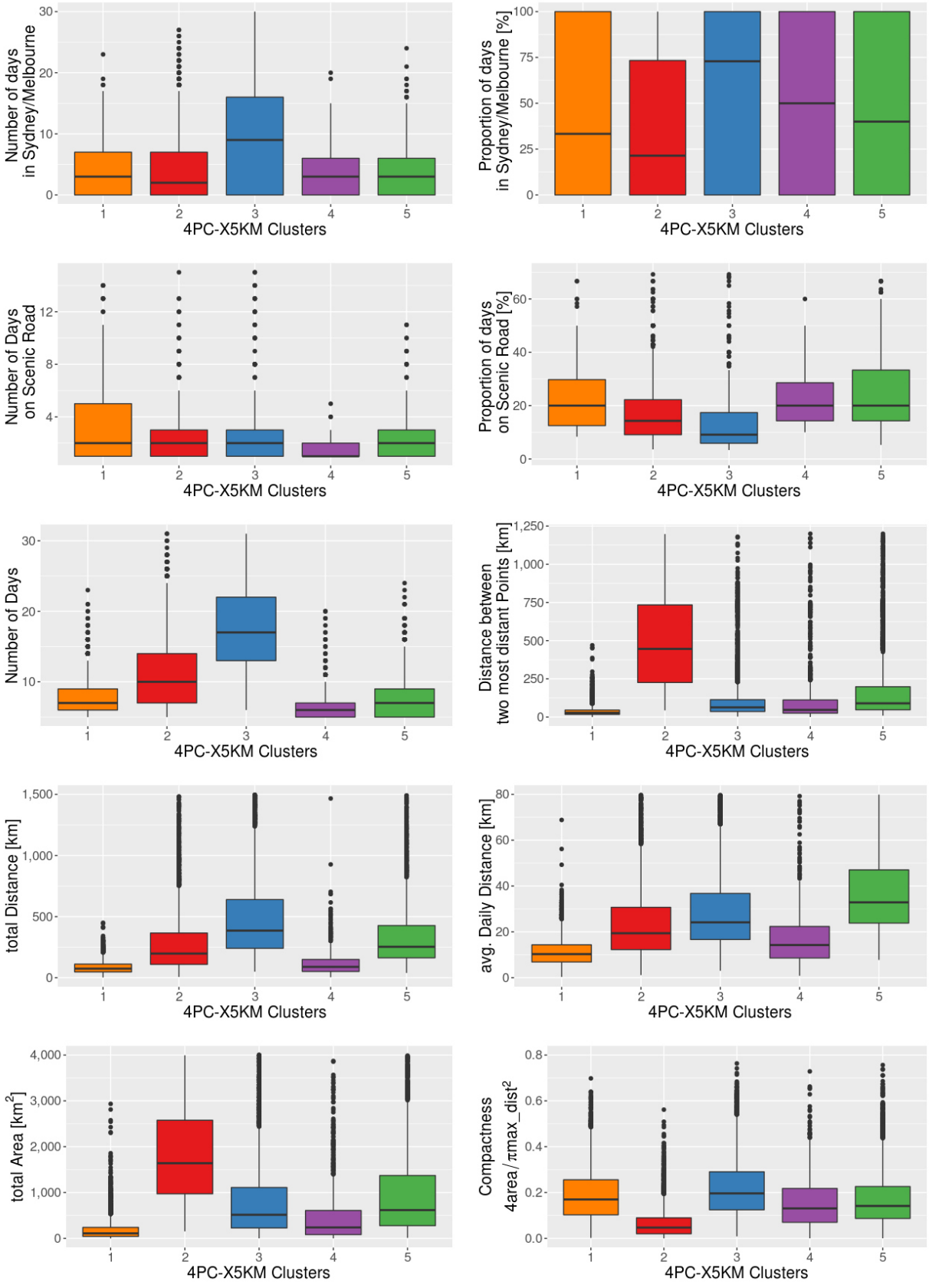

Figure 4.16: Boxplots for eight different original variables across the five clusters of the 4PC-X5KM approach. Part 1

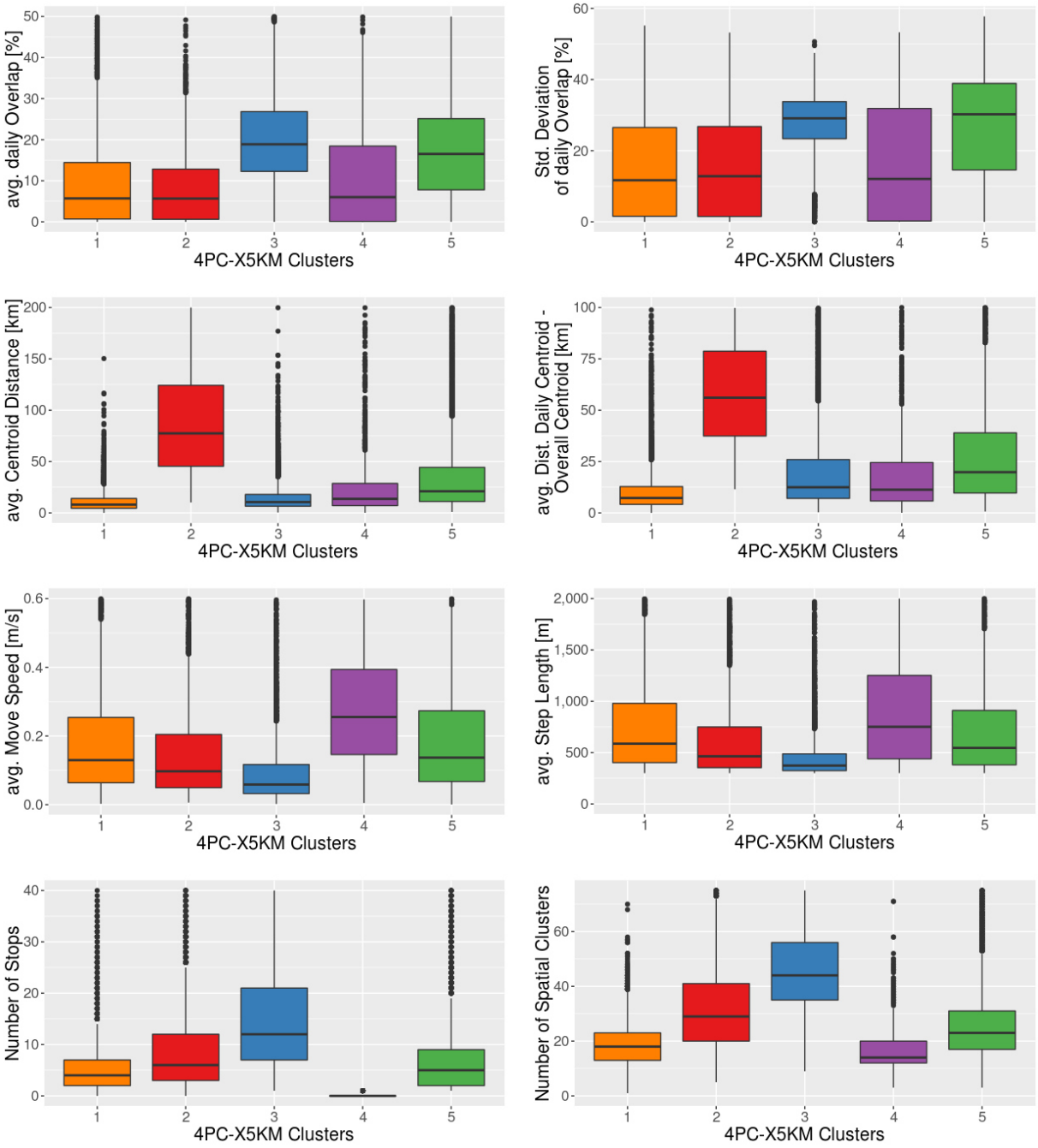

Figure 4.17: Boxplots for seven different original variables across the five clusters of the 4PC-X5KM approach. Part 2

5. Temporal & Spatial Analysis of User Types in Cities

The last few sections have led to the division of the individual users into clusters and later to user types, based on their spatio-temporal footprints. By doing this, we can answer the first research question stated in **section 1.3**.

The goal of this thesis, however, is not only to divide users into groups, but also to get a deeper understanding of both the temporal and spatial characteristics of these worked-out user types, as stated in research question 2:

Research Question 2: *What are the spatio-temporal usage patterns of the identified types of users in the two cities of Melbourne and Sydney? Can individual areas be characterized based on temporal usage patterns of different user types?*

Due to the building of user types based on the results of the last few sections (interpretation found in **section 4.5**), we can now speak of user types instead of clusters. Accordingly, we analyze the SSE patterns of the user types in the two cities of Sydney and Melbourne to answer research question 2. In **section 5.1**, the temporal characteristics of the formed user types are investigated, followed by an investigation in the spatial characteristics of the user types in the two cities of Melbourne and Sydney in **section 5.2**. In these two sections, we solely analyze the findings of the visualizations, whereas a discussion of the found patterns follows in the discussion **section 6**.

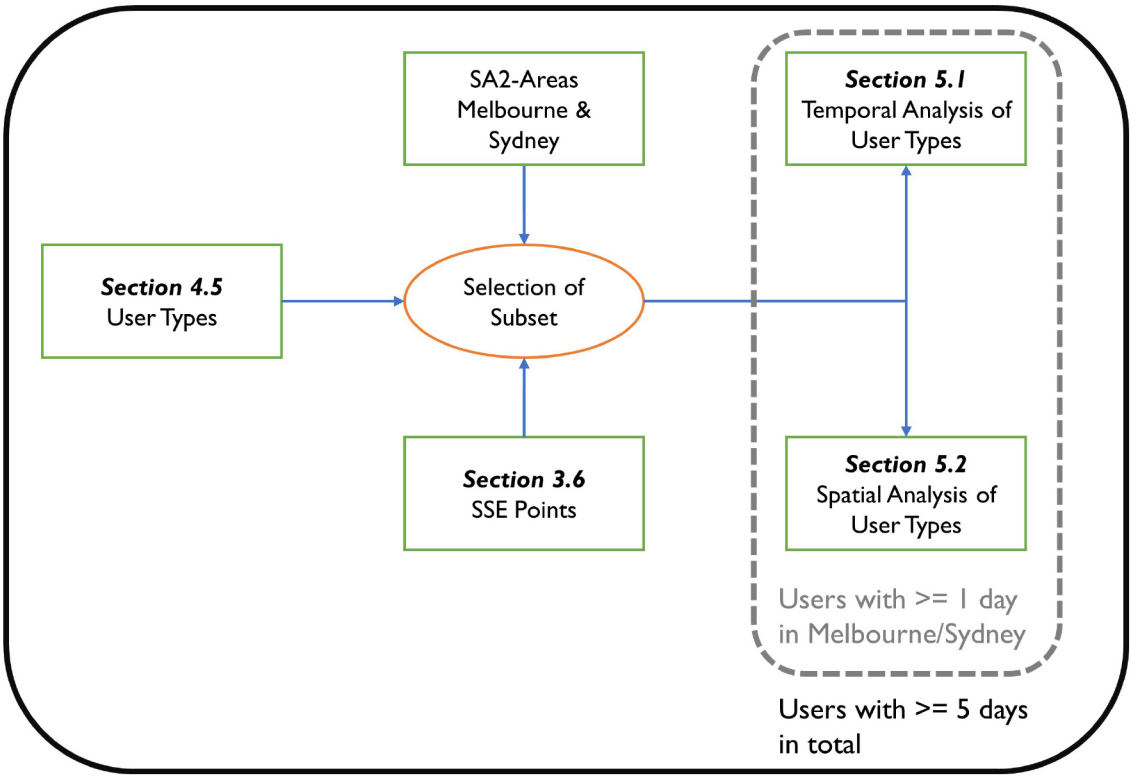

Figure 5.1: Workflow of Chapter 5

5.1 Temporal Characteristics of User Types

In a first step, we analyze the temporal characteristics of the different user types. Three different approaches were chosen to confirm or reject the second hypothesis:

Hypothesis 2: *Different user type use the two investigated cities in different ways. They visit different places and have different temporal usage patterns. Accordingly, different urban areas show distinct visiting patterns by different user types.*

5.1.1 Daily Temporal Distribution

Method

First, we have a look at the daily temporal distribution of SSE points to see whether there are differences among the different user types. We therefore aggregate all SSE points per city and user type for the whole month to one single day and analyze the differences between the different user types. We both use the absolute temporal distribution of SSE points and the normalized distribution for each user type to find distinct patterns for each user type.

SSE Points are used since they serve as a proxy for certain actions the individual users do, i.e. they either start, stop, or end a session at said point. These points are the easiest

way to investigate patterns, since they are independent from the individual movement patterns, but still reflect the spatio-temporal whereabouts of the individual users belonging to a user type.

Results

The users of type C (the "Commuter") combined show by far the largest number of SSE points per time of day (hourly interval) in Sydney and Melbourne, when looking at the absolute hourly distribution (top plots in Figure 5.2). The numbers of the other two user types, T ("Overland Delivery Driver"/ The "Tourist") and E (The "Excursionist") then do not differ much.

When looking at the pattern itself, Figure 5.2 shows that user type C differs a lot from the other two user types. The differences between the maximum and the minimum values seems much bigger for user type C. This can be confirmed by looking at the standard deviations in Table 5.1, which shows that type C's standard deviation is much higher than the ones of type T or E. The difference between the two other clusters however is not that large, also seen in the rather close standard deviations.

The time series for each user type and city further differ when looking at the normalized time series (bottom plots in Figure 5.2). In Melbourne, the pattern for user types T and E do not differ much from each other. An actual morning peak cannot be seen, but rather an increasing curve of activity until noon. After a short decrease until 2 pm, a daily peak can then be perceived at 3 pm (type T), respectively 4 pm (type E). User type C 's pattern however is different regarding two aspects. First, a morning (9 am), noon and afternoon peak (4 – 5 pm) is shown. Second, the shown values between 10 am and 4 pm (working hours) are much smaller than the ones for the two other user types T and E.

In Sydney, similar patterns and differences can be seen, however, not as strong as the ones for Melbourne. Again, user type C shows three peaks, whereas the two other user types T and E only show one peak (3 – 5 pm) with a continuous increase during the day. During the working hours, the differences between the user types are not as strong as for Melbourne, showing only small variations among the different user types.

A further interesting insight is that user type C's pattern in the two city differs from each other. Whereas it shows an equally high morning and an afternoon peak in Sydney, cluster C shows a much higher afternoon peak in Melbourne.

Table 5.1: Hourly mean and standard deviation values for the three user types in the two cities

	User type T		User type C		User type E	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Melbourne	1274.41	896.93	7314.00	4130.08	1612.42	974.05
Sydney	1166.21	791.09	6743.46	3917.24	1426.13	837.23

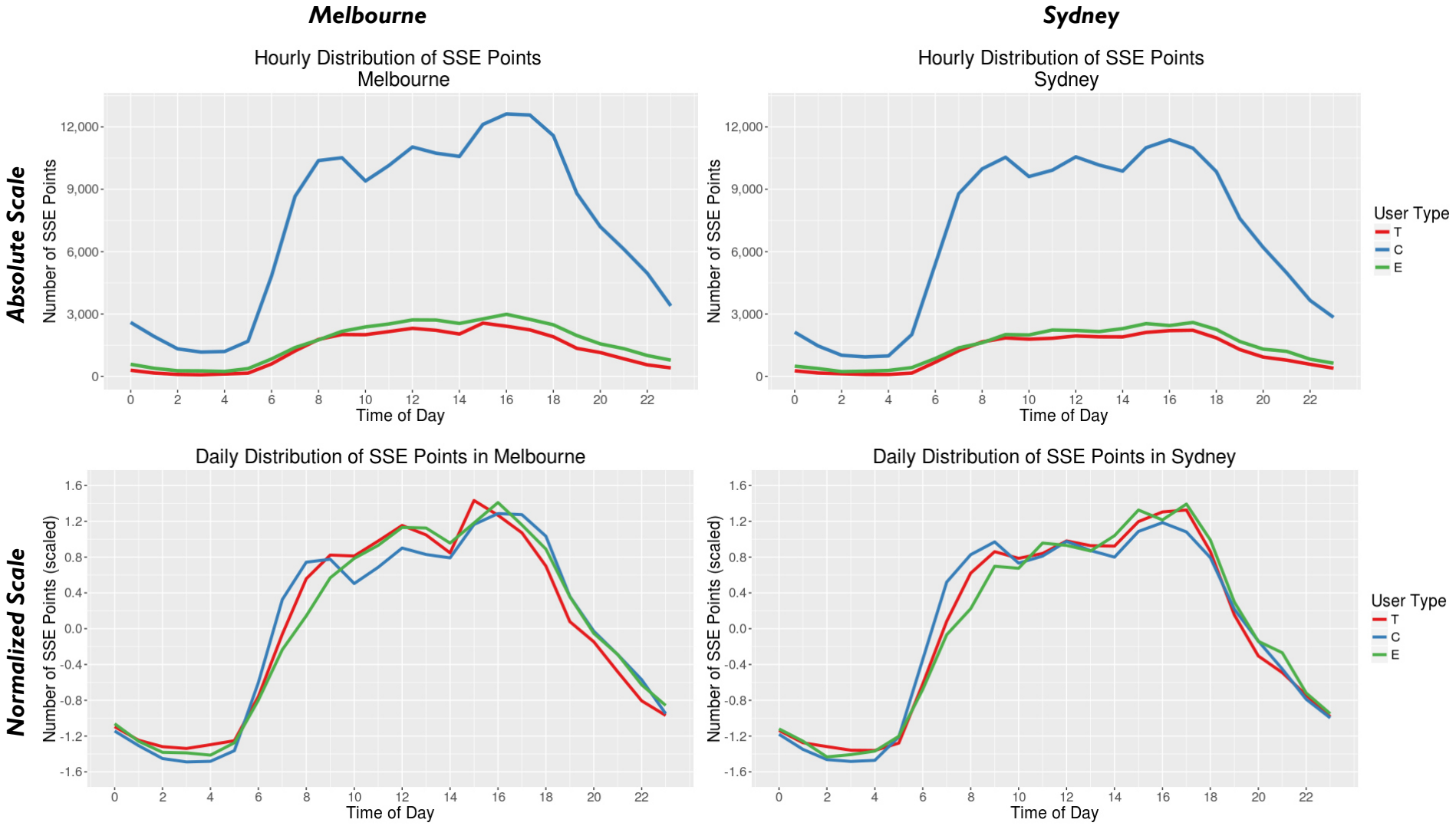

Figure 5.2: Absolute scale (top) and normalized scale of (bottom) daily distribution of the SSE points for user types T, C and E for Melbourne (left) and Sydney (right)

5.1.2 Weekly Temporal Distribution

Method

In this section, we take a similar approach as in the last **section 5.1.1**, this time however using an aggregated weekly distribution. To do that, data for the first 28 days (4 weeks) are taken and aggregated into two-hour windows over the course of a week and to create a composite view of a typical week. By only taking data for the first 28 days, no weekday is overrepresented as each one occurs exactly four times.

This analysis is both done with absolute values (number of SSE points in said time interval, top plots in Figure 5.3) as well with normalized values (bottom plots Figure 5.3). We therefore scaled the absolute values of the different user types to get time series that show the same variance. This allows us to find patterns that could have been undetected due to extreme outliers in the absolute numbers.

Results

Similar to **section 5.1.1**, much higher values are shown for user type C than for user type T or E. In addition to that, the standard deviation for user type C is much higher both for the daily as well as the hourly pattern (Table 5.2). The higher daily standard deviation is especially shown in the plots with absolute numbers (top plots in Figure 5.3). Here, the differences between the maximum and the minimum values is much higher for type C than for the other two user types. User types T and E's pattern does not differ much from each other. Both have about similar mean and standard deviation values, whereby the ones for user type E is a bit higher.

Looking at the peaks on weekdays, it can again be seen that type C has a much higher afternoon peak than a morning peak in Melbourne, whereas in Sydney the two peaks are about equally high. On weekends, two daily peaks can only be seen on Saturday in Melbourne. On Saturday in Sydney as well as on Sundays in both cities, only one peak for user type C can be seen, having its largest values between noon and 4 pm. For the two other user types, much more stable daily pattern can be seen, where the differences between weekdays and weekends are not as big as the ones for user type C. This is also confirmed by the values found in Table 5.2, which show that user type C has a much higher daily standard deviation. The aggregated pattern for type E shows that on weekends, there is a higher peak in the morning than in the afternoon, whereas on weekdays the pattern is reversed.

The time series for each user type and city additionally differ when looking at the scaled values (bottom plots in Figure 5.3). Whereas the three user types have similar amplitudes for weekdays, they differ on weekends. There, magnitudes for user type C are much smaller than the ones of user type T and E. User type E even has the highest amplitude for Saturday morning when looking at the Sydney graph. We can further state that type E users tend to use the app more in the afternoon than in the morning, except on weekends where the amplitudes for the morning are higher.

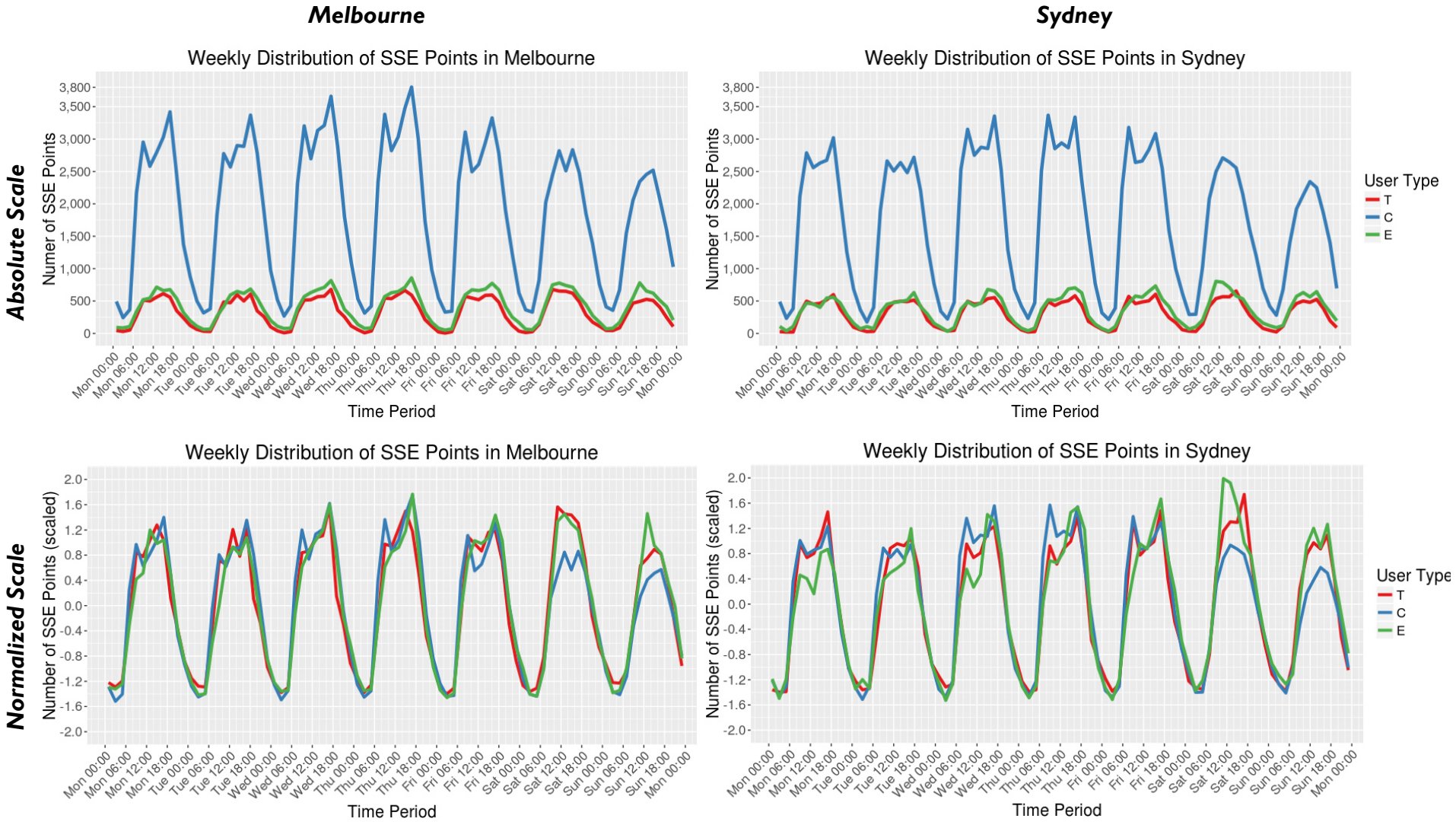

Figure 5.3: Absolute scale (top) and normalized scale of (bottom) weekly distribution of the SSE points for user types T, C and E for Melbourne (left) and Sydney (right)

Table 5.2: Hourly and daily mean and standard deviation values for the three user types in the two cities

		User type T		User type C		User type E	
		Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Melbourne	Hourly	323.35	227.23	1895.36	1089.02	415.43	249.71
	Daily	3880.14	284.43	22744.29	2792.68	4985.14	328.58
Sydney	Hourly	303.06	202.30	1741.80	1035.43	367.25	219.33
	Daily	3636.71	238.96	20901.57	2607.68	4407.00	396.87

5.1.3 Periodicity

Method

In a third temporal analysis step, we are interested in the cyclic processes, i.e. the degree of periodicity found in the SSE points of the individual user types. According to Ahas et al. (2015), cyclic processes in spatio-temporal data are more likely short-term (24-hour, weekly, seasonal cycle) than long-term and occur with a certain regularity. They can be seen in commuting, tourism, seasonal employment and agriculture in the case of climate zones (Panda et al. 2002; Silm & Ahas 2010 in Ahas et al. 2015). According to Filion (2000) and Ahas et al. (2015) such cyclic patterns can be in a next step used to detect monofunctional places in cities, i.e. areas and places in a city that are only used by a group of people in a short period during the day.

Several methods exist to compute the degree of periodicity for a given frequency. Here, we use Fast Fourier Transformation (FFT), a method proposed by Calabrese et al. (2010). FFT is a signal processing technique that uses sine and cosine functions to compute a magnitude, reflecting the degree of periodicity for a certain time interval (Han et al. 2012). The highest magnitude then indicates the frequency with the highest periodicity in the given data.

In our case, FFT is applied on the scaled time series for each user type and city. The resulting set of magnitudes is then again scaled to better compare the individual user types with each other. By doing that, we can determine differences among user types and can confirm or discard the findings of the visual interpretation of the weekly distribution analysis, found in section 5.1.2.

Results

For Melbourne (top plot in Figure 5.7), the highest magnitudes are given for the daily circle. Accordingly, daily patterns repeat with a much higher magnitude than other patterns. The second highest patterns can be found for 12 hour intervals and weekly intervals. Likewise, the same pattern can be seen for Sydney (bottom plot in Figure 5.7).

The differences between the user types in both cases, however, confirm the patterns seen in Figure 5.6. Type C users (the "Commuter") show a smaller daily periodicity than type T and E users. This can be explained due to the smaller amplitudes of user type C during the weekends compared to the weekdays, seen in Figure 5.5 and Figure 5.6. Type T and E users, however, show a much more stable behavior (i.e. amplitudes) during all days of

the week. Accordingly, this results in a higher amplitude in for the daily cycle. Although having smaller daily cycle values, type T users have a comparably higher amplitude for the weekly cycle.

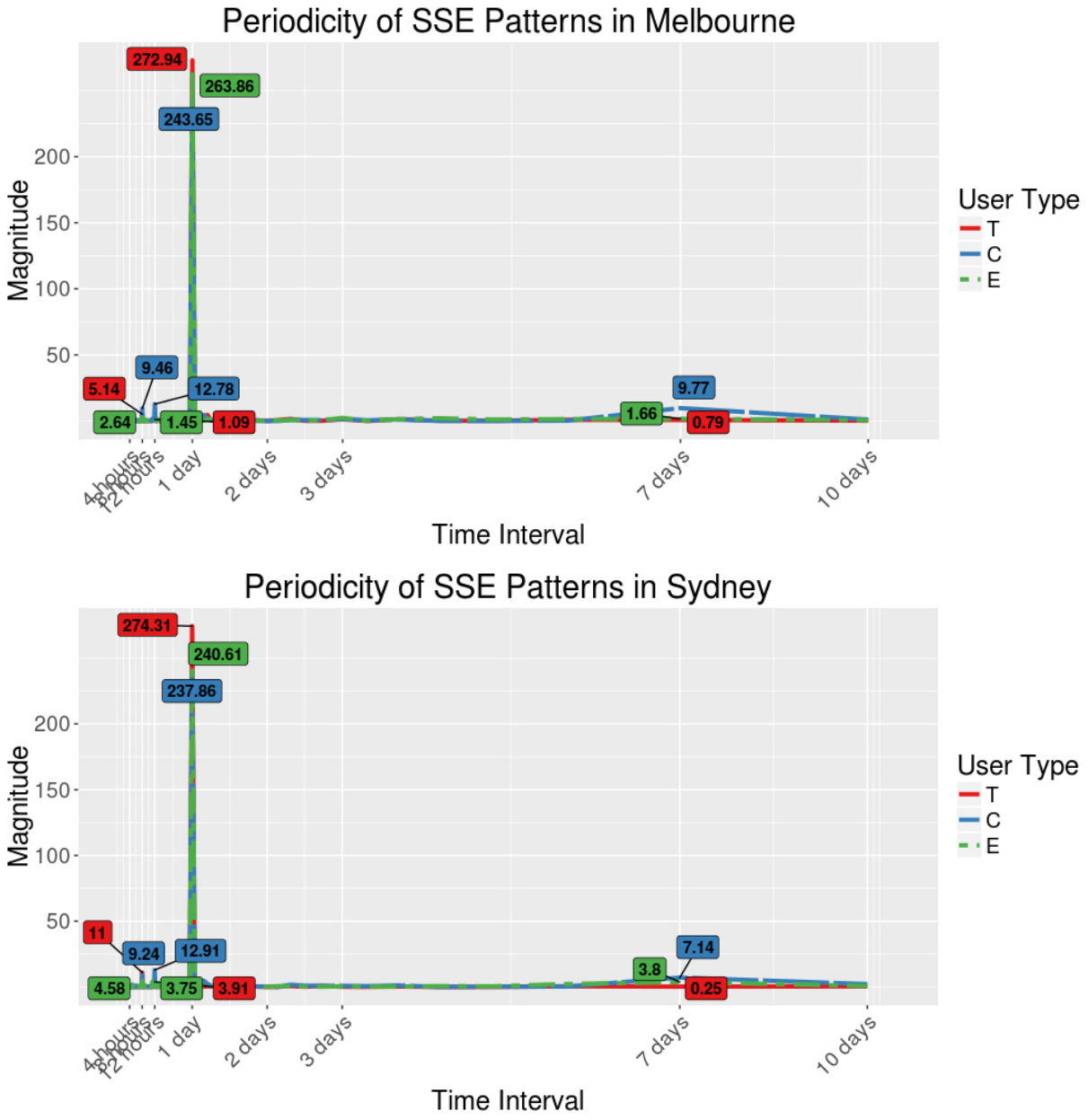

Figure 5.4: Fourier Transformation of the scaled and aggregated time series for Melbourne (top) and Sydney (bottom). Marked are values for 8, 12 and 24 hours as well as 180 hours (approximately 1 week).

5.2 Spatial Characteristics of the User Types

In order to answer research question 2, we further carried out an investigation of the spatial characteristics of the different user types. To this end, five different approaches involving the SSE points have been chosen: absolute distribution, relative distribution, location quotient, global spatial autocorrelation, and connectivity. The goal is to determine SA2 areas of Melbourne and Sydney that show a distinct visiting pattern for a certain user type that differs distinctly from the patterns other SA2 areas.

A visual exploration of the overall areas of the Greater Melbourne Area and the Greater Sydney Area (see section 3.4.1) have shown to be too big for a thorough qualitative analysis and interpretation. We therefore decided to only investigate a spatial subset of the Metropolitan areas, the individual SA2 areas that contain at least 12 points per square kilometer. Several values have been tested and visualized. 12 SSE points/km² has then been shown to generate a reasonable number of remaining areas that are, furthermore, bundled around the city centers.

5.2.1 Spatial Distribution of SSE Points per User Type

Method

The spatial characteristics of the different user types will be analyzed in a first step, by looking at the absolute spatial distribution of SSE points by user type, in each city. Since the SA2 areas have different sizes, we aggregate the number of SSE points per square kilometer for each user type. To reduce the influence of very active users, multiple SSE points of the same user within a certain area were removed. The used approach presents us with a better way to compare both user types and areas respectively than by just looking at the absolute numbers of SSE points per area.

Comments on Visualizations

In Figure 5.8 and Figure 5.9, each SA2 area is colored per number of SSE points per square kilometer using a sequential color scheme from *ColorBrewer* (Neuwirth 2014). Using a different boundary color, we further highlighted areas that show a distinct pattern both in this analysis and in the analyses of the next sections. For Melbourne, the highlighted areas are the City Center with its central business district (CBD, *C*), Melbourne Airport (*A*), and St Kilda (*S*), Melbourne's "*favorite beachside suburb*" (Visit Victoria 2016). In Sydney, the highlighted areas are Sydney Airport (*A*), the City Center (*C*, including *"The Rocks"* and CBD) and two of Sydney's most famous beach areas (Destination NSW 2017): Bondi Beach (*B*) and Manly Beach (*M*).

Results

The most obvious observation is that the more central a SA2 area, the higher the number of SSE points per square kilometer. The areas remaining based on the spatial subset (>= 12 SSE points/km²) are bundled around the city center and its central business district (highlighted with a "C") in the case of Melbourne. Likewise, this is given for Sydney, here, however, also SA2 areas along the northern coast are remaining.

Figure 5.8 and Figure 5.9 show that user type C has by far the biggest amount of SSE points per area of all user types. User type T and E on the other hand do not differ much from each other, both in the number of SSE points/km² and the presented SA2 areas. Based on the patterns found in the visualizations, it must be clarified that a visualization of the absolute distribution does not bare much information that can be used to generate additional information for the user types. Accordingly, more thorough investigation with different approaches is needed to investigate the different point patterns.

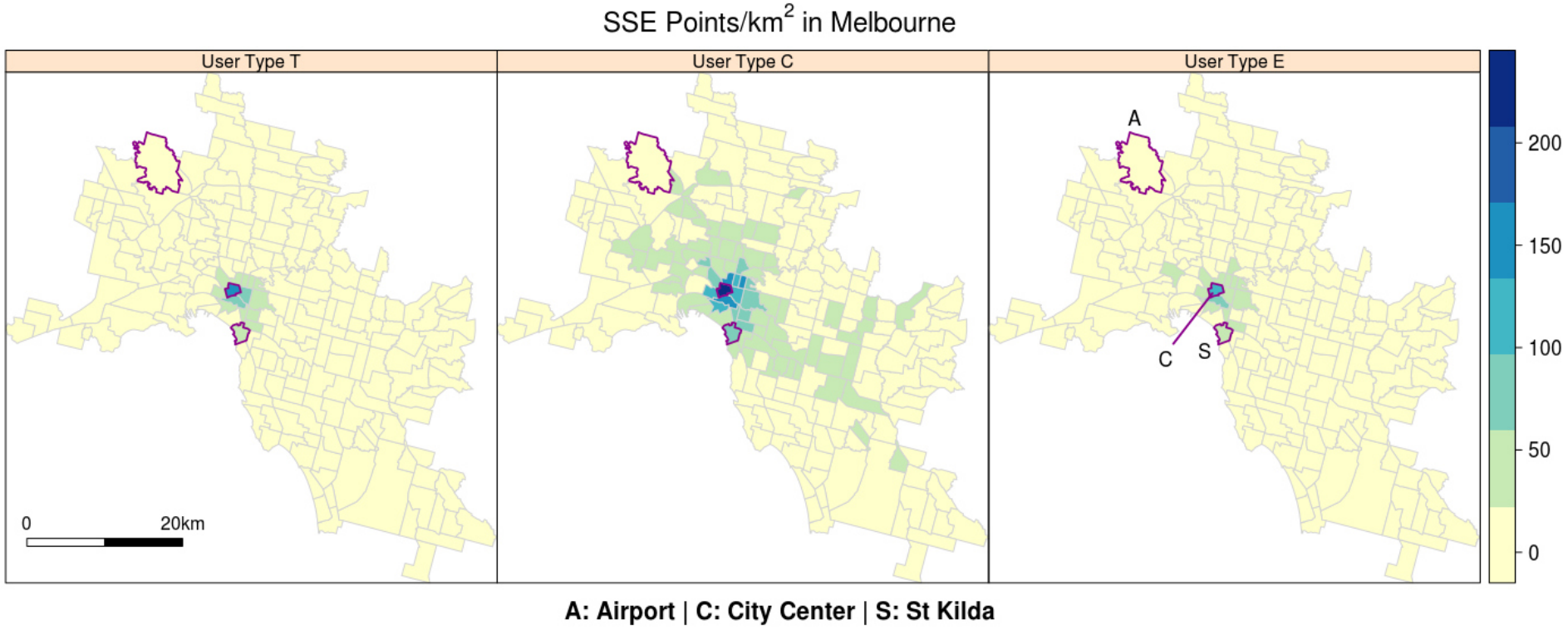

Figure 5.5: Number of SSE points per square kilometer for each user type and SA2 area of Melbourne, where number of SSE points/km^2 > 12.

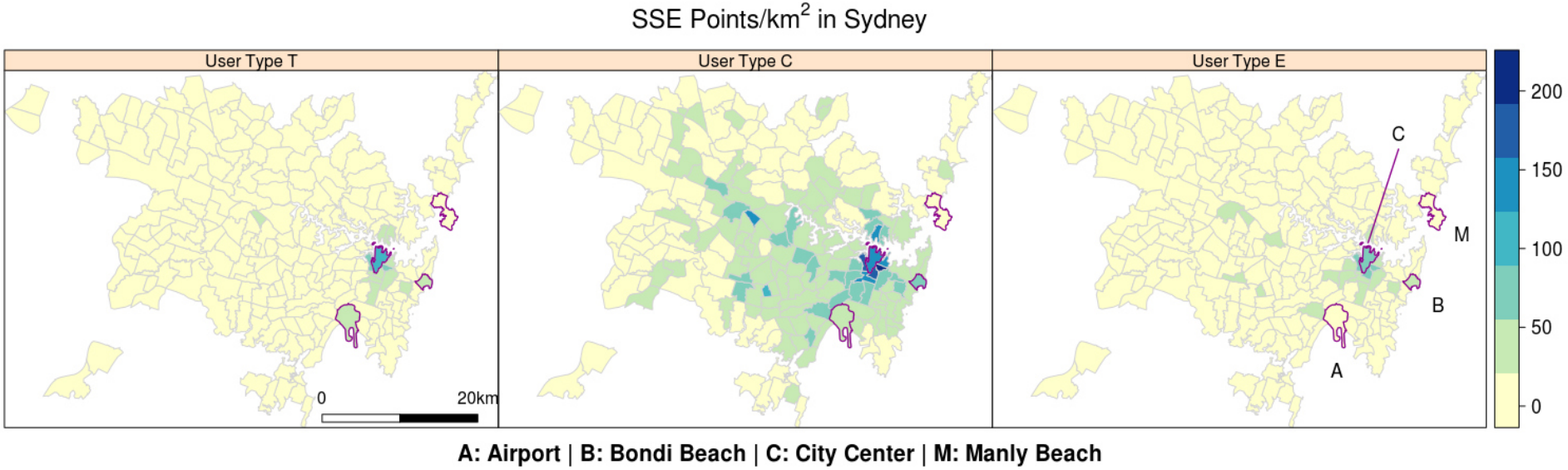

Figure 5.6: Number of SSE points per square kilometer for each user type and SA2 area of Sydney, where number of SSE points/km² > 12.

5.2.2 Relative Distribution – Percentage of User Type per Area

Methods

Besides the absolute distribution of the SSE points per user type (**section 5.2.1**), we are additionally interested in the relative distribution. For each SA2 area in both cities, we determine the proportion of SSE points belonging to each user type. Consequently, we get a percentage value for each SA2 area and user type, whereas a high percentage value for a certain user type reflects a high proportion of said user type in that area. Using a visual analysis of the relative distribution can therefore lead to the identification of SA2 areas where certain user types are disproportionately present. High values, i.e. darker colors in the choropleth maps in Figure 5.10 and Figure 5.11 indicate that in these areas, a given user type is relatively overrepresented.

Results Melbourne

In the case of Melbourne in Figure 5.10, we can again see that user type C is the most representative user in all areas, i.e. shows the highest percentage values. User type T and E have similar percentage values, whereas user type E has a slightly smaller standard deviation in its values.

The spatial patterns of the different user types manifest interesting patterns. The values of user type T (the "Overland Driver"/the "Tourist") are high in the highlighted SA2 areas, i.e. the Airport, the City Center, St Kilda and down the coast (Tullamarine, Essendon, Avalon). SSE points belonging to user type C (the "Commuter") are relatively high in the areas around the city center and rather low in the highlighted areas. A special pattern can be seen in the SSE points of user type E. Here, we can see that the percentage values are higher the farther away the areas are from the city center, especially in the western SA2 areas. The pattern however seems much more homogeneously distributed than the ones for the other two user types, which is also reflected by the smaller variance in the percentage values.

Results Sydney

In the case of Sydney (Figure 5.11), we can observe a structurally similar situation as in the Melbourne. Again, user type C has the highest values in all areas as well as the highest variance. For user type T, again the magenta-highlighted areas (Airport, City Center, Bondi Beach, and Manly Beach) show high percentage values. Low percentage values can be seen in the areas farther away from the city center, especially in the west.

For user type C, rather small values are shown in the City Center and the Airport Area. For the two other highlighted areas, Bondi Beach and Manly Beach, smaller values can be seen, but not as small as the ones for Airport and City Center. Rather higher values can especially be seen in the areas south and west of the City Center. User type E shows a similar image as in the case of Melbourne, with a rather small variance and a lot of areas with values around the mean. High values can be seen for areas further away from the city center and the other highlighted areas. Lower values can be seen around the highlighted areas, but not in the areas themselves.

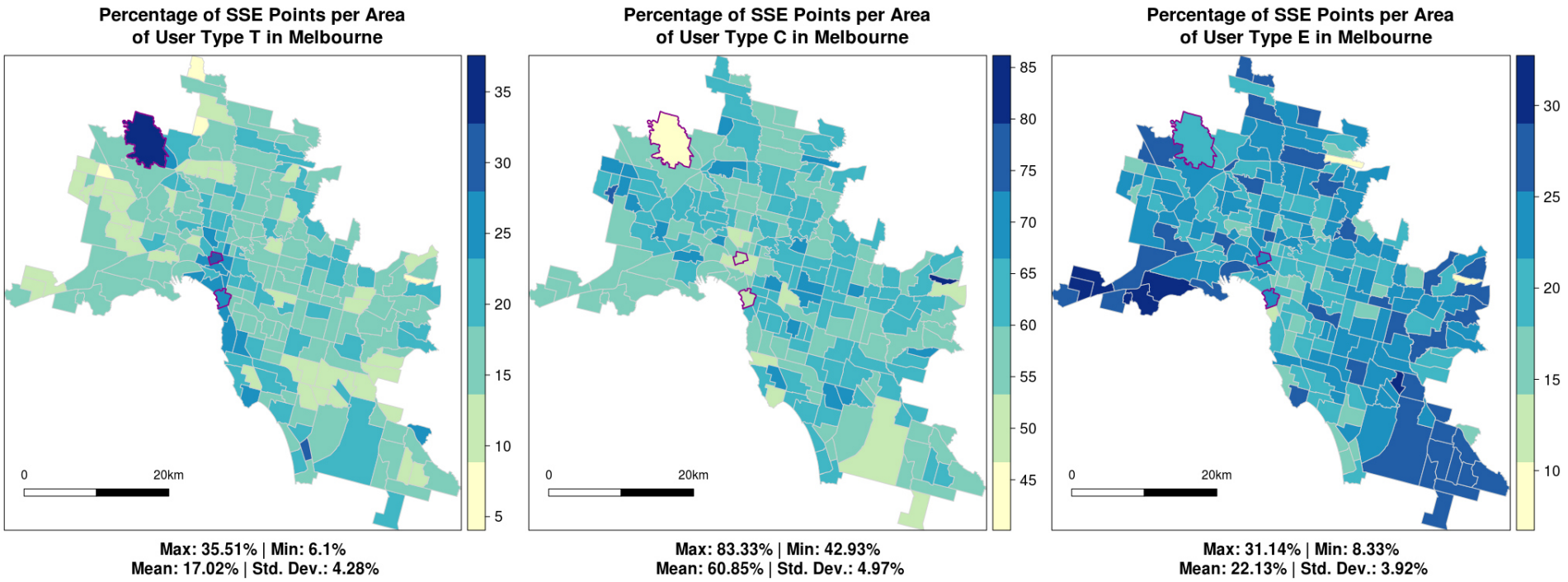

Figure 5.7: Percentage of SSE points belonging to a user type of each area for Melbourne

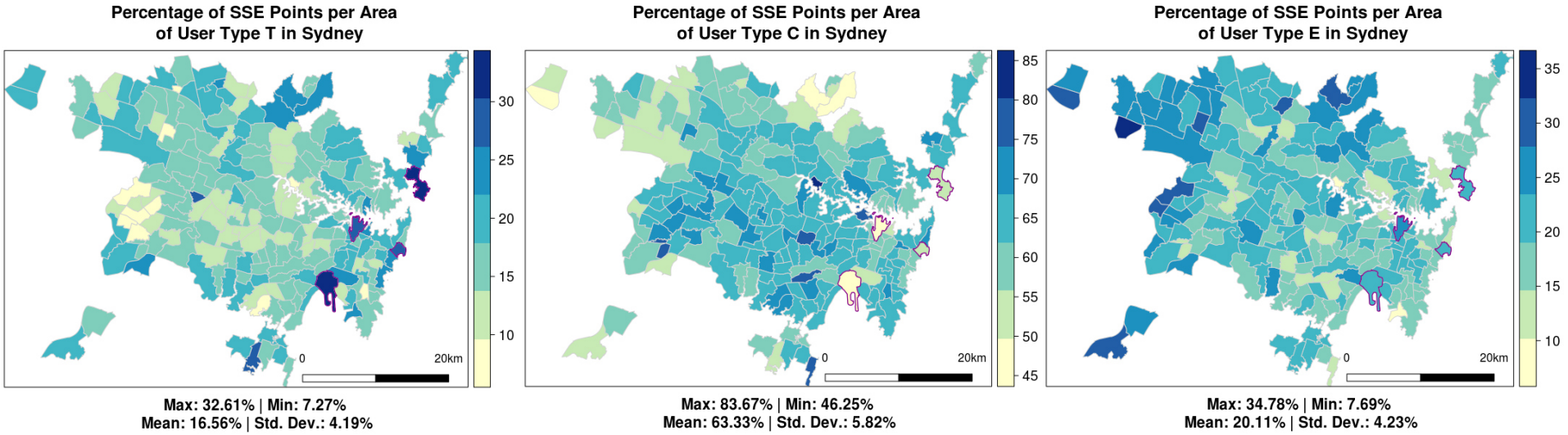

Figure 5.8: Percentage of SSE points belonging to a user type of each area for Sydney

5.2.3 Location Quotient

Methods

In the next step, we are interested in finding SA2 areas for each user type that show a non-standard visiting pattern, i.e. finding SA2 areas with a high or low *location quotient*. The location quotient compares the ratio of a local density of a phenomenon *i* in an area *p* to the overall density of that phenomenon in a reference area (in our case the whole cities) (Reades et al. 2009; Jiang et al. 2015). Accordingly, a high location quotient is given for areas in which a certain user type is overrepresented, whereas a small location quotient is given for areas in which the user type is underrepresented.

To compute the location quotient, we compute for each area and user types the relative difference between the observed value and the expected value. In that case, the observed value is the computed proportion of SSE points belonging to a user type in a certain area, i.e. the percentage values from **section 5.2.2**. The expected value however is the mean value of all proportions of said user type over the whole city. The relative difference in percent, i.e. the location quotient is therefore defined as

$$location\ quotient = \frac{observed\ value - expected\ value}{observed\ value} * 100$$

.

As an example: if 50 percent of the SSE points found in area A belong to user type X, but the overall mean value of said user type is only 30 percent, the location quotient in that area and for that user type is +40 percent.

Comments on Visualizations

In Figure 5.12 and Figure 5.13, the location quotients are visualized using a diverging color scheme from ColorBrewer (Neuwirth 2014). On the bottom of each map there is an additional histogram showing the distribution of the location quotient values for each user type. To compare the individual visualizations and user types in both cities, we have chosen the same color scheme breaks for all visualizations. These breaks [-18,-6,6,18] have been chosen manually based on an investigation of the individual histograms.

In the following two sections, we use the terms overrepresentation and underrepresentation when referring to areas that show values above the two highest breaks in the color schme (-18%, 18%). Over-/underrepresentation does not necessarily mean that given user type has the highest amount of SSE points in that area, but has a high relative difference to the expected mean value.

In Figure 5.14, we have assigned each SA2 area in both cities to the user type that has the highest positive location quotient. Due to that we need a qualitative color scheme and therefore use the same colors for each user type as in the boxplots in **section 4.5.5**, also depicted in Table 4.7. By using this approach, we are directly able to show which areas are typical for a certain user type to visit, based on the location quotient.

Results Melbourne

One of the first things that can be realized when looking at both the maps and the histograms in Figure 5.12, is that far more areas are shown with values below -18% or above 18% for the user types T (left) and E (right) than for user type C (middle). Accordingly, this means that in many more areas, user types T and E are over- or underrepresented. Opposite to that is user type C which is much more homogeneously distributed over the whole city, additionally depicted in the histogram with a lot of values bundled around zero. Furthermore, both user types T and E show far smaller negative relative differences than user type C, with minimum values of up to -180% in comparison to user type C with a minimum of only -42%. These areas can therefore be thought of as areas that are rather less visited than other areas.

Looking at the individual user type and its over/-underrepresented area, it shows that user type T is overrepresented in the airport as well as areas around the center and south of St Kilda (highlighted). On the other hand, user type T is underrepresented in areas further away from the city center. User type C shows a complete opposite pattern and is underrepresented in areas in which user type T is overrepresented. Especially in the Airport area as well as the city center and its southern surroundings, user type C is underrepresented. Most of the remaining other areas do not show big under-/overrepresentations, accordingly, it can be stated that user type C has a more equal spatially distribution of their SSE points.

User type E (bottom left) is represented about average at the highlighted areas (Airport, St Kilda, and City Center). On the contrary, it is especially underrepresented in the areas around St Kilda and down the coastline. On the other hand, user type E has a relative overrepresentation in the areas in the south east and in the west.

The top map in Figure 5.14 further shows that the areas around the City Center, along the Coast and the Airport are assigned to user type T. The other areas are then assigned to user type C and E, whereas the areas further away from the center are more likely assigned to user type E.

Results Sydney

In Figure 5.13, a similar image than in Figure 5.12 is shown; again, user type T (left) and E (right) show more areas in which they are either under- or overrepresented than user type C (middle). This is shown in the histogram which depicts that user type T's value are bundled around zero without a large number of outliers.Furthermore, user type C has again smaller extreme values [-37,25] than the two other user types (up to -162%). This confirms the finding for Melbourne, that user type C users are more equally distributed in space than the two other user types.

The left map for user type T shows that, like in the case of Melbourne, user type T is overrepresented in the City Center, the Airport as well as the areas around St Kilda and Bondi Beach, i.e. the highlighted areas. User type C has, as mentioned before, only a few areas in which it is underrepresented. These areas are again the ones in which user type T is overrepresented. User type E then has a lot of areas in which it is overrepresented, however, these areas lay at the outskirts of the city in the south, west, and north.

The map on the bottom of Figure 5.14 shows again each area assigned to the user type, which has the highest relative positive difference between the observed and the expected value. Again, the areas that are the farthest away from the city center are assigned to user type E, whereas the city center and the beaches are assigned to user type T. The areas assigned to user type C are mainly around the harbor and the city center.

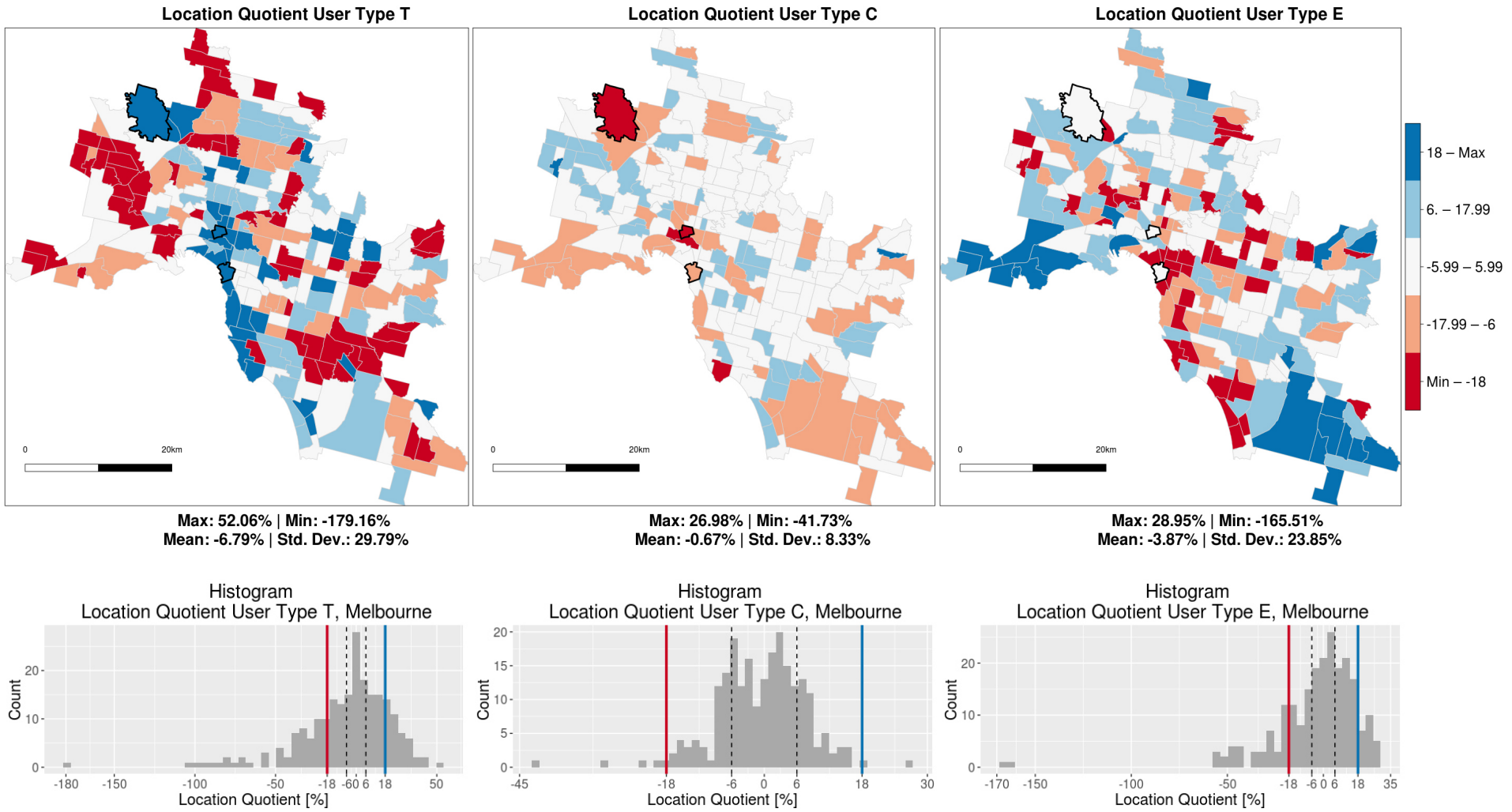

Figure 5.9: Computed location quotients for each SA2 area and user type in Melbourne (top) and histograms of location quotients (bottom)

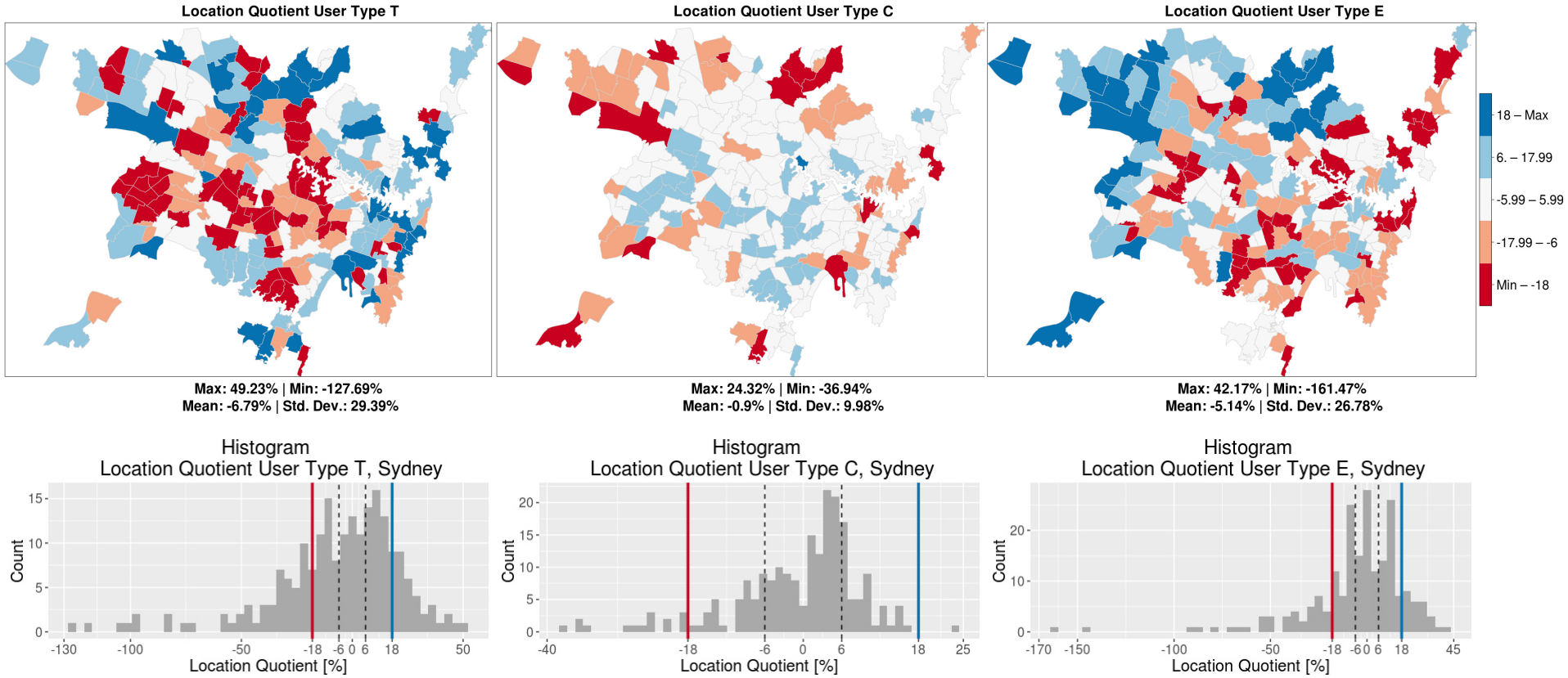

Figure 5.10: Computed location quotients for each SA2 area and user type in Sydney (top) and histograms of location quotients (bottom)

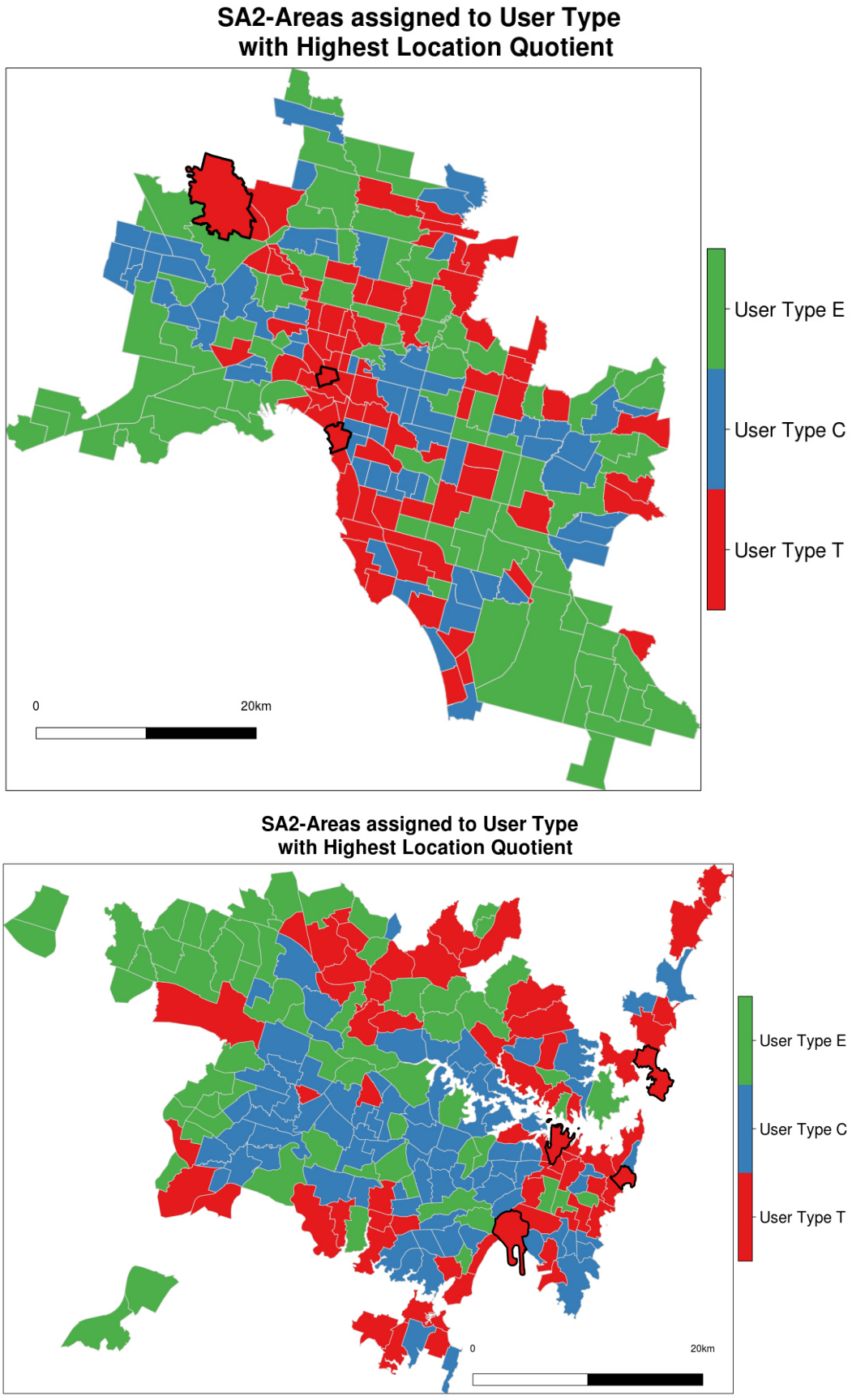

Figure 5.11: Each SA2-area assigned to user type with the highest location quotient (top: Melbourne, bottom: Sydney)

5.2.4 Global Spatial Autocorrelation

Methods

We further test the relative distribution of SSE points for spatial autocorrelation to test whether the found patterns per user type are randomly generated or not. Spatial autocorrelation describes the tendency of nearby areas to have similar values of a given variable, based on their attributes (Brunsdon & Comber 2015). In this case, spatial autocorrelation tests can be used to understand the degree to which the percentage values per SA2 area and user type are similar to other nearby areas. We are interested in the global pattern, hence compute the global spatial autocorrelation for each user type and city. The most popular indicator to measure the spatial autocorrelation is the Moran's I coefficient, which ranges from -1 to +1, with negative values reflecting negative autocorrelation and positive values positive spatial autocorrelation. A value near zero indicates a random spatial pattern.

We compute Moran's I value as well as its corresponding p-values for the relative distribution values of the user types for both cities. The computed values can then be compared for each user type in each city, but not across the two cities (Esri 2016). The null hypothesis for this test is that the values for all spatial objects are randomly distributed, meaning that no spatial autocorrelation is given and Moran's I is equal to zero.

A simulation based approach was chosen to compute the Moran's I value and the p-value, as suggested by Brunsdon & Comber (2015). With this approach, a certain number of random permutations of the data are taken and assign to the individual areas. For each random permutation as well as the actual distribution of the values, Moran's I values are then computed. If the null hypothesis is true, then the "*probability of drawing the observed data is the same as any other permutation*" of the data among the areas (Brunsdon & Comber 2015, p.234).

Results

Based on the p-values in Table 5.3, there is strong evidence to reject the null hypothesis and to accept the alternative, i.e. that the percentage values shown in Figure 5.10 and Figure 5.11 are spatially autocorrelated. More importantly, the test shows that the individual patterns are not randomly generated.

The Moran's I values further show that all user types in both cities show positive autocorrelations. For Melbourne, the highest Moran's I value is given for user type T and the lowest for user type C. For Sydney, user type E presents the highest Moran's I value, whereas user type T shows the smallest Moran's I.

Table 5.3: Moran's I-values and p-values for the relative spatial distribution of the individual user types the two cities

	User type T		User type C		User type E	
	Moran's I	P-Value	Moran's I	P-Value	Moran's I	P-Value
Melbourne	0.339	< 2.2e-16	0.192	1.244e-06	0.240	2.548e-09
Sydney	0.325	4.665e-12	0.352	9.478e-14	0.402	< 2.2e-16

5.2.5 Connectivity of SA2 Areas

Methods

An additional approach to analyze the characteristics of the different user types is by analyzing the degree of connectivity between individual SA2 areas for each user type. We therefore establish an origin-destination matrix (ODM), with the vertex being the individual SA2 areas and the edges being the number of trajectories that connect these SA2 area. In our case, we generate the ODM based on all SSE points of all sessions. For each session, we tested whether an SSE point as well as the next SSE point lies within the city. If both were within the city borders, the resulting connection was stored in a new table and further aggregated with the other connections to form the ODM. We therefore only counted distinct OD trips per users across time in order to remove the bias of dominance of a single user in the data set, which would blur the connectivity intensity.

A visualization of the ODM for each user type and city then shows the areas with the highest degree of connectivity to other areas. Using that approach, we can show that the user type influences not only the overall spatial and temporal patterns, but also the magnitude and order in which certain areas are visited.

Comment on Visualizations

In Figure 5.15 and Figure 5.17, the origin-destination matrices for the different user types in the two cities are visualized. In Figure 5.16 and Figure 5.18, zoomed-in visualizations of the ODM around the city center for the two most diverse user types T and C are shown. In all visualizations, both the line width as well as the line color are based on the intensity of the connection between two areas. The connections with the lowest values in the ODMs were removed to get a better interpretable visualization. Accordingly, we can both make points about the intensity of the different connections as well as the degree of connectivity of the different areas per user type.

The values in the legend show the number of moves (movement between two consecutive SSE points) between two areas. Furthermore, in all figures, the same areas are again emphasized, this time with a blue fill color to delineate it from the yellow and red lines. For Melbourne, these are the Airport, the City Center, and St Kilda; for Sydney, the Airport, City Center, Bondi Beach, and Manly Beach.

On the bottom of each map there is an additional histogram showing the distribution of the connection values for each user type. Highlighted with a red line is the chosen threshold for each user type that separates the in the map shown connections from the not-shown connections.

Results Melbourne

The three plots in Figure 5.15, as well as the zoomed plots in Figure 5.16 show that user type C (middle) has by far the biggest connectivity between individual areas. Trips within the city are therefore much more frequent than for other user types. Not only more areas are connected with each other, also the absolute number of trips, the intensity between them is much higher than for other user types. Rather special is that for all user

types, the highest connectivity values are between Coburg and North Coburg and vice-versa, followed by the connection between the City Center and Southbank south of it. An additional area that has high connectivity values in all user type is Dandenong South in the south-east.

The visualization of the ODM of user type T (left) shows that user type T only has a few areas that are connected to each other. Moreover, these areas are rather around the City Center than in the suburbs. What can further be seen is that again, user type T has higher values in the Airport and the St Kilda areas then in other areas. Some degree of connectivity is also shown towards areas where freeways exit Melbourne (west, north-east, and south-east). The ODM visualization of user type E (right) shows something in-between the visualizations of user types T and C. Here, we have much more areas that are connected with each other than user type T, but not as much as user type E. Accordingly, that pattern is rather difficult to interpret.

The histograms shown on the bottom of Figure 5.15 confirm the described patterns. Although the highest value in all histograms is given for 0, meaning that most areas are not connected with each other, differences between the distributions can be seen. First, much more areas are not connected to each other for user types T and E than for user type C. Second, when neglecting the 0 values, the histograms of user types T and E show a strong skewed right distribution, whereas the distribution for user type C is much more unimodal.

Results Sydney

At first sight, Figure 5.17 does not show differences between the different user types as big as in the case of Melbourne (Figure 5.15). What can be seen is that user type C (middle) and E (right) have much more connections between the different areas than user type T (left). Furthermore it can be seen that certain axes can be seen that have been built based on the connections of user type T. These axes run along the main routes that lead into/exit the city center. These axes can additionally be seen in the two other user types; however, they are not formed as strongly as for user type T, also due to the much denser pattern. Overall, it can be stated that for user type C and E, the pattern of the connections is much more diverse. Furthermore, much more high-valued connections can be seen for user type C and E. The connections for user type T are distributed much sparser around certain main axes and centers (City Center, Bondi, Manly and Airport).

The findings above are confirmed when zooming in into the SA2 areas around the city center (Figure 5.18) and comparing user type T (left) and user type C (right). Around the city center, fewer areas are connected to each other for user type T than for user type C. Whereas the intensities of the individual connections do not differ much between the two user types, user type T has a few areas with a higher intensity of connections (City Center, Bondi, Manly and Airport).

The histograms shown on the bottom of Figure 5.17 again confirm the described pattern. Again, the highest values are given for zero, meaning that most areas are not connected with each other. Furthermore, the distribution of user type C can be interpreted as much more unimodal than the ones of user types T and E, when neglecting the zero values.

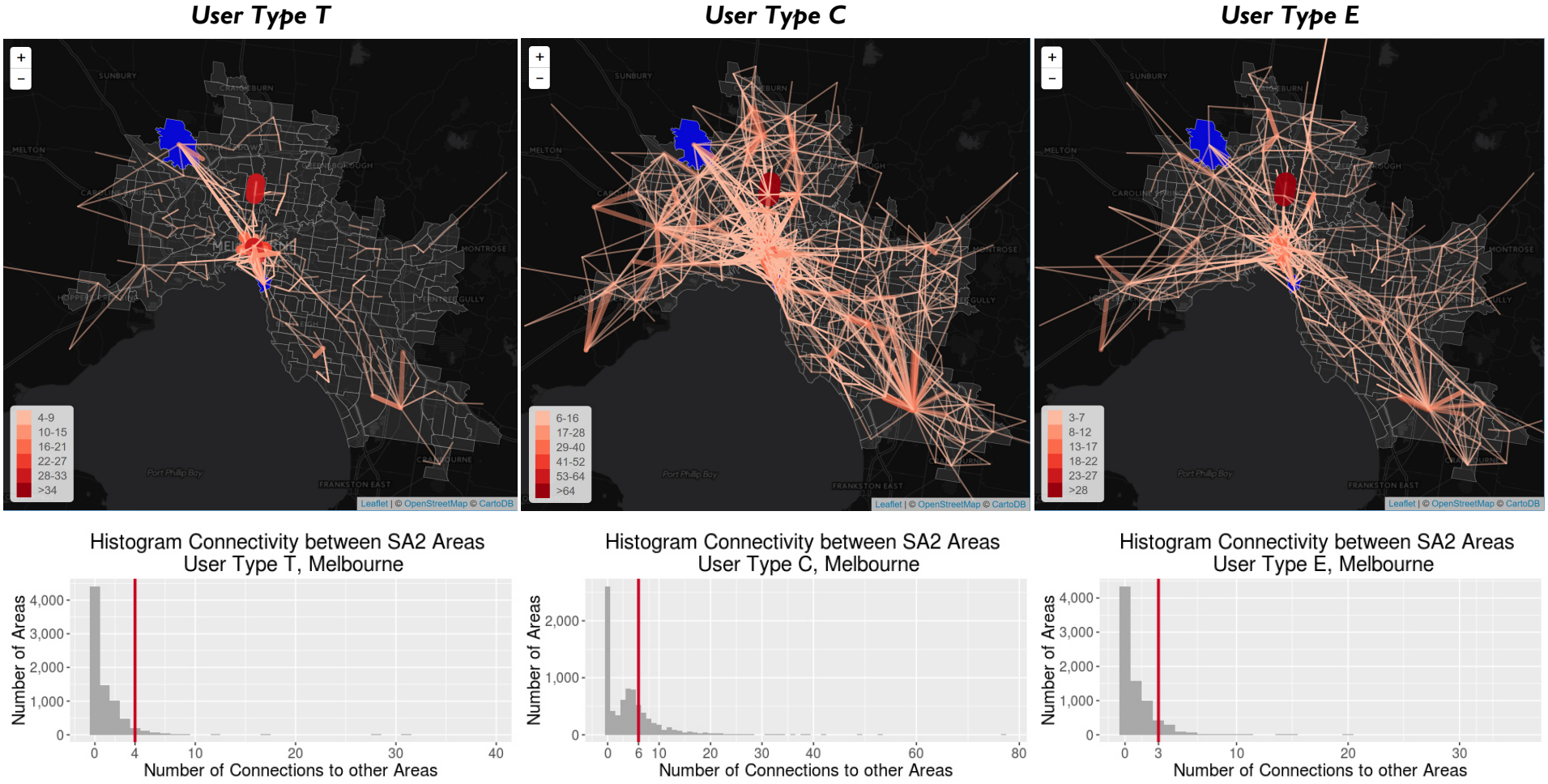

Figure 5.12: Top: Visualized Origin-Destination matrices for Melbourne, highlighted in blue are again the City Center, St Kilda, and the Airport. Bottom: Respective histograms of connectivity values.

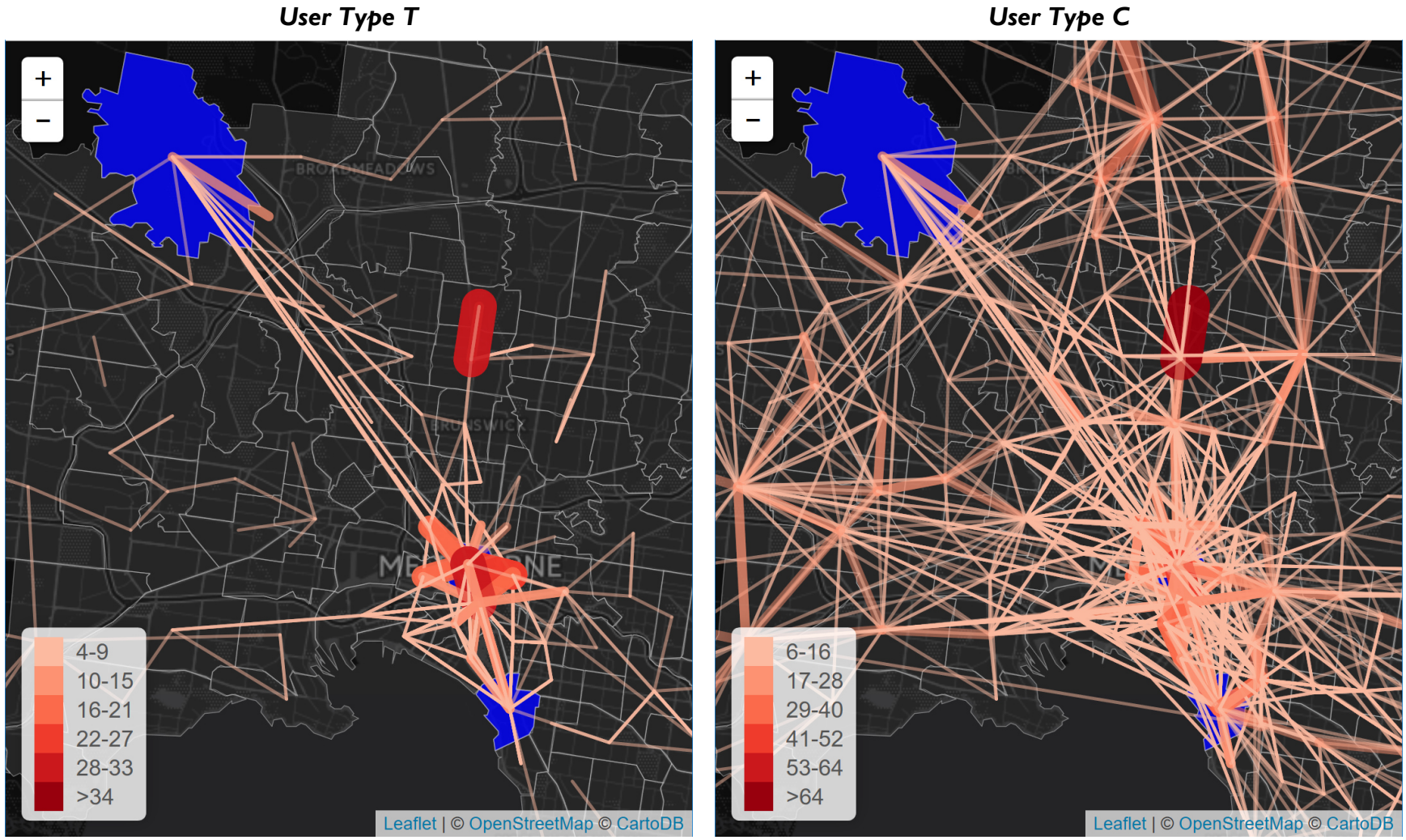

Figure 5.13: Visualized Origin-Destination Matrices for Melbourne, zoomed in. User type T (left), User type C (right). Highlighted in blue are again the City Center ("The Rocks"), Bondi Beach, Manly Beach, and the Airport.

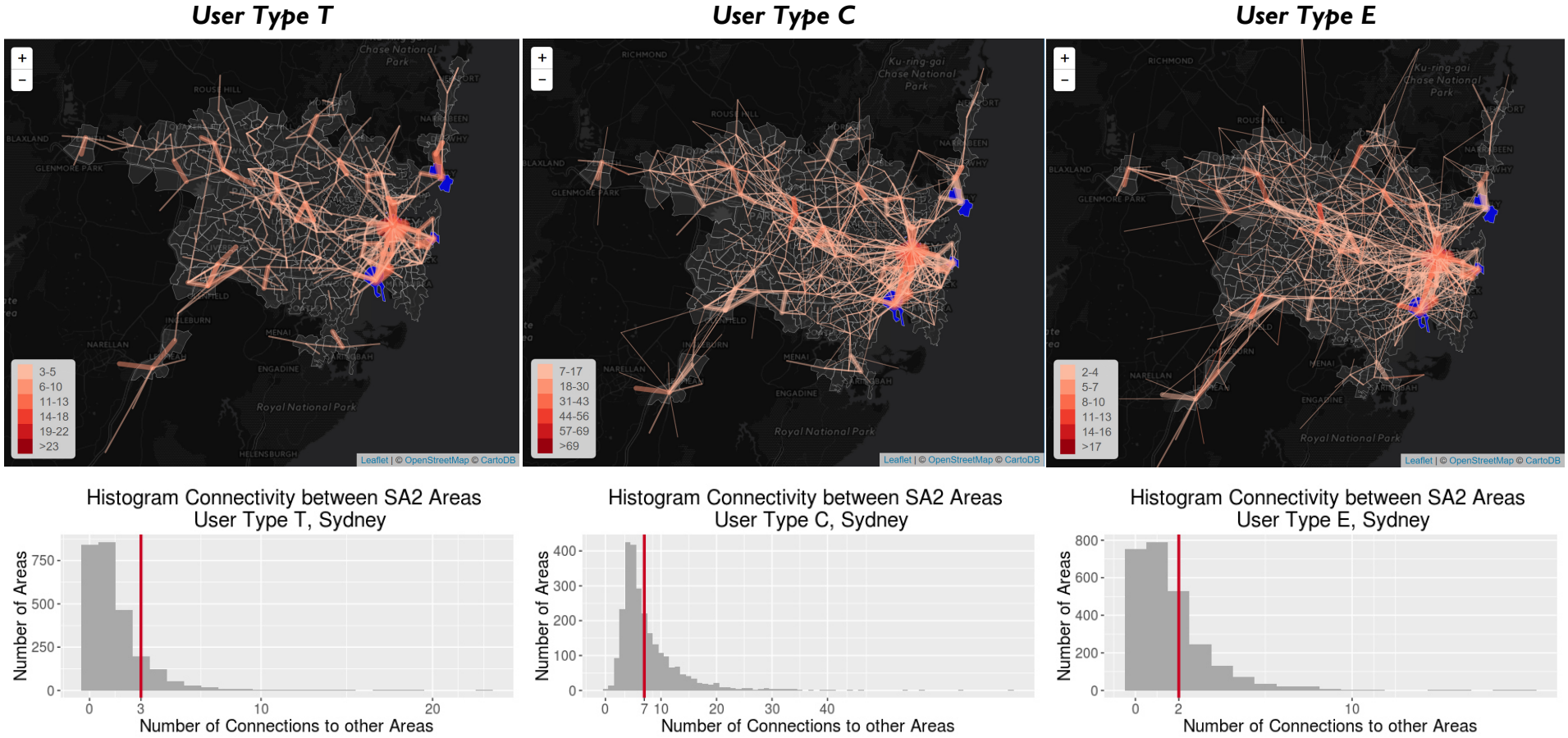

Figure 5.14: Top: Visualized Origin-Destination matrices for Sydney, highlighted in blue are again the City Center ("The Rocks"), Bondi Beach, Manly Beach, and the Airport. Bottom: Respective histograms of connectivity values.

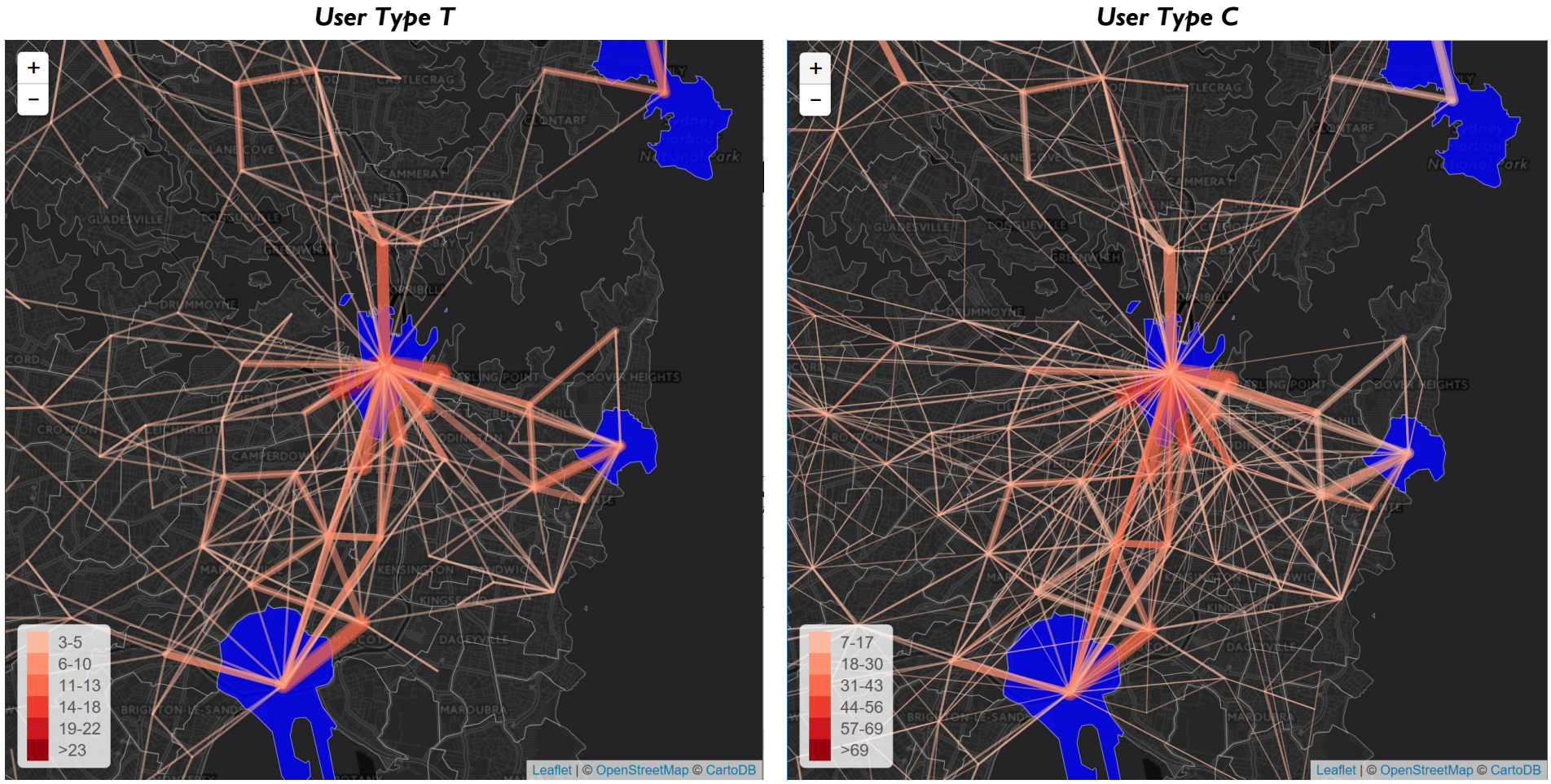

Figure 5.15: Visualized Origin-Destination Matrices for Sydney, zoomed in. User type T (left), User type C (right). Highlighted in blue are again the City Center ("The Rocks"), Bondi Beach, Manly Beach, and the Airport.

6. Discussion

In this chapter, we discuss the results of the two previous sections, namely the characterization of users into user types (section 4) as well as the temporal and spatial analysis of the user types found in the two cities of Melbourne and Sydney (section 5). We further link the results back to the research questions, as stated in section 1.3.

In a first part, section 6.1 discusses the methodology used to compute the individual user types (RQ1). In section 6.2.1, the temporal characteristics of the user types found will be revised. The spatial characteristics of the user types in the two cities will then be discussed in section 6.2.2 (RQ2).

6.1 User Characterization in the Absence of Ground Truth

Characterizing users from a large set of navigation data requires several pre-processing steps (as described in Chapter 3) and the application of a series of un-supervised learning methods (Chapter 4). In contrast to several other studies, including the tourist study by Shoval & Isaacson (2007) or the mobility study by Trasarti et al. (2011), there is no additional knowledge about the nature of the individual users found in the data. It is unknown where individual users live and, more importantly, it is unknown why the users use the app. No statements can be made about the users' intentions, i.e. whether they use the app for daily commuting or during their holidays. Accordingly, we lack ground truth of the people's characteristics using the app.

Based on our assumption that different users will favor distinct areas of the cities (and thus have different spatio-temporal footprints), we assume that different user types generate different spatio-temporal footprints. These, as well as further assumptions, have served as a basis for the first research question:

Research Question 1: *How can individual users of the navigation app be characterized based on their spatio-temporal footprints in the absence of ground truth? What are the principal factors describing the different user types?*

Spatio-Temporal Footprints

To divide the users into groups based on unsupervised methods, we assume that different users show a variety of spatio-temporal footprints. The spatio-temporal footprints as defined in **section 3.1.1** are the sum of the movement recorded for a given user, independent of the actual locations the movement is recorded at. Accordingly, a user living in Melbourne commuting to work daily and a user in Sydney doing the same may have the same spatio-temporal footprint, although living in two different places. A user on a road trip who is also visiting the two cities, however, shows a completely different image and therefore has a different spatio-temporal footprint.

In the case presented, we went a step further than previous studies and did not analyze the trajectories itself, but the actual spatio-temporal footprints the individual users possess. By doing so, we also uncover an additional temporal component and we can analyze changes in a user's movement patterns over the course of several days. This way, we may have lost some information regarding the trajectories themselves, but gain a more detailed user profile.

Certain computed measures of the spatio-temporal footprints such as the overall area, are very sensitive to outliers. Since several additional measures are based on the areas (daily or overall), the chosen approach is dependent on a good implementation of the function used to compute the area, in our case the concave hull. Nevertheless, we have tried to overcome these problems beforehand, by removing flawed sessions in our data cleaning step (**section 3.7**). To further remove the influence of individual, undetected and flawed sessions, we have only computed values about the daily and overall patterns.

Assumption of Road-Bound Movement

Ranacher et al. (2016) argue that GPS data originating from cars are effected by measurement errors leading to the positioning of GPS tuples off the road network. Furthermore, they argue that measurement errors lead to the overestimation of the distance measurements themselves. Map-matching is therefore recommended when using GPS data of cars.

In our case, a first visual inspection of some of the unprocessed trajectories showed that most tested trajectories followed the road network. Accordingly, no map-matching techniques were necessary, since we assumed that most of the non-tested trajectories would follow the road network.

Due to the omission of the map-matching step, the removal of both spatial and temporal outliers was carried out using a different approach. We computed the temporal threshold, distance, velocity, and acceleration between two consecutive GPS tuples and used them to remove spatio-temporal outliers.

Segmentation of Trajectories into Moves and SSE Points

In this thesis, we built the database following the ontology (see section 3.5) presented by Renso et al. (2012). Said ontology divides the individual trajectories into move segments and stop points. In the case of Renso et al.'s (2012) study however, they were presented with permanently sampled data without temporal gaps with no data. Due to that, the ontology used here was extended by additionally using the start and end points of the individual sessions, forming the SSE points.

For the actual stop detection within the sessions, an SQL-algorithm was created that used both a spatial and a temporal threshold to test each GPS tuple for stopping behavior. As a temporal threshold, we decided to use a period of five minutes. We argued that within a stop lasting at least five minutes, the respective user can interact with the environment and the actual location. Accordingly, a found stop segment can be depicted as a minimal period required for accomplishing a pragmatic task. By neglecting stops that last less than five minutes, we ensure that no stops are detected that may be due to congestion or measuring errors. Nevertheless, we must be aware that choosing different temporal and spatial thresholds, the resulting SSE point patterns as found in the results section 5.2 would look completely different.

Principal Component Analysis and Clustering

The approach to apply principal component analysis on the spatio-temporal footprints first and then do clustering on a set of principle components was chosen due to several reasons. Firstly, by choosing only a limited amount of principle components, the influence of noise is minimized which can lead to statistical benefits for the clustering approach (Bro & Smilde 2014). Secondly, the principle components present us additional underlying patterns, that can be found by the dimensionality reduction (Bro & Smilde 2003).

The ranking of the different clustering approaches has lead us to two rankings in which four of the five best approaches in each ranking coincided. On the one hand, the consensus of the two rankings supports the approach chosen. On the other hand, the resulting clustering algorithm should, besides the good statistical properties, also be interpretable (James et al. 2013). The two approaches (3PC-X3KM and 4PC-X5KM) with the best statistical properties have shown that this is an important step, since the second but best approach was better interpretable than the first ranked approach.

The resulting interpretation of the ultimately chosen approach (4PC-X5KM) has then been carried out based on boxplots of the original values of the spatio-temporal footprint. This interpretation then has led to the formation of three user types. Accordingly, only three of the original five clusters were then taken into the spatio-temporal analysis of the SSE points. This approach can be criticized, since 5'530 (28.9%) of the 19'106 originally used users for the PCA were removed from the further analysis. The users removed, however, showed only very small numbers in terms of the spatio-temporal footprints. An interpretation of these clusters therefore proved to be difficult. Furthermore, the analysis of the SSE points then showed that the neglected two clusters are only responsible for less than 10% of the SSE points within the two cities. We argue that

due to the small usage numbers, the SSE point pattern of the two neglected clusters would distort the overall point pattern in both space and time.

User Type Interpretation

The three user types presented in **section 4.5** were elaborated based on the patterns found in the boxplot and the following interpretations. Thus, we tried to bring the individual clusters into a human-interpretable form in order to assign them some sort of label. The chosen labels have been discussed with various researchers, however, we must be aware that these labels were given based on the assumptions made by the author of this thesis. Various studies from different fields were consulted, but no other possible user type descriptions could be found. Due to that, the actual labels must be treated with caution, since a validation proves to be difficult without the respective ground truth. The analysis of the SSE point patterns in **sections 5.1** and **5.2**, however, has shown that the chosen labels might reflect a pattern that is imaginable for the three user types.

Summary and Outlook

To conclude and to answer the first research question, we have shown that different user types can be found based on spatio-temporal footprints and in the absence of ground truth. Thus, the validation of the individual methods leading to the different clusters has been carried by the different cluster validation methods as well as the rank aggregation. The validation of the description of the user types found that is based on the elaborated clusters has, however, proven to be rather difficult without the corresponding ground truth. It would therefore be interesting, if a similar approach was carried out on similar data with a corresponding ground truth. It would then be easier to validate the results of the clustering as well as the interpretation itself.

6.2 Assessment of Temporal and Spatial Characteristics of User Types

In the previous sections, the first research question and the formation of the different user types has been discussed in detail. The question following RQ1 is how the actual user types use different areas of cities as addressed in the research question 2:

Research Question 2:	*What are the spatio-temporal usage patterns of the identified types of users in the two cities of Melbourne and Sydney? Can individual areas be characterized based on temporal usage patterns of different user types?*

Following is a discussion of the temporal characteristics of the user types found in **section 6.2.1** and a discussion of the spatial characteristics of the identified user types in **section 6.2.2**.

6.2.1 Temporal Characteristics

Assessment of the Daily and Weekly Patterns

Both the daily as well as the weekly patterns (sections 5.1 and 5.1.2) show that, in all time periods, most of the SSE points within the two cities belong to user type C (commuter type). This is unsurprising, when we compare this number to the absolute number of users within the two cities belonging to type C (3'321, see Table 4.7 in section 4.5.5). This number is higher than the numbers of users of type T (2'140) or E (2'572). Not only has user type C the biggest number of users within the city, it also has the highest percentage of SSE points that lie within the two cities' borders (49.85%, type T: 32.76%, type E: 42.46%). These findings can therefore serve as a confirmation of the clustering result and its interpretation, namely that users belonging to type C are indeed more active in cities than other user types.

Another interesting finding presented in Table 4.7, is the fact, 64.01% of all the users belonging to user type T (touristic type) have at least spent one day in Melbourne or Sydney. This stands in contrast to user types C and E, of which only 56.66% and 58.83% of the users have spent a day in the cities. This finding stands, however, in great contrast to the boxplot showing the proportion of days spent in one of the two cities (top left boxplot in Figure 4.16). There, it is shown that user type T has a far smaller median and IQR than user types C and E. Based on that, we argue that users of type T overall spend less days in the cities than other user types, but on the contrary, have a higher likelihood that they visit a city. Accordingly, this reflects our user type interpretation, namely that users of type T visit a lot of places in a short amount of time due to their tourist behavior.

The normalized weekly pattern in section 5.1.2 further shows that the degree of activity of type C is not as constant over both the course of a week and the course of a day as for other clusters. Again, this may serve as a confirmation of the assumption that users of type C live in or around the city, using the app mostly for commuting between home and work. Due to that, smaller numbers and different patterns can be seen on weekends, when users belonging to type C are believed not to have to drive to work.

The two other user types, T and E, manifest a contrasting pattern to user type C. They have a far more stable pattern when looking at the weekly distribution. Especially users belonging to type T (touristic) are not dependent on working hours and therefore use the app on a more constant basis over both the course of a day or week. Regarding the numbers, users of type E (excursionist) show a stable pattern when comparing weekdays and weekends. The pattern shown, however, is rather different, having the peak values in the afternoon on weekdays and in the mornings on weekends. This might indicate the nature of users belonging to type E, namely to carry out excursionist short trips in the evenings of weekdays or in the mornings on weekends, which require the assistance of a navigation aid.

Differences in Daily/Weekly Patterns between Groups

The weekly SSE point pattern of the two cities of Sydney and Melbourne as well as of the different user types is different. This has been anticipated and confirms our findings, as

various previous studies have shown comparable results using other data sets. Kling & Pozdnoukhov (2012) worked with Twitter and Foursquare data and showed that different types of social media topics have different daily and weekly patterns with different peaks, similar to our user types.

Similar results are presented by Grauwin et al. (2015) which used several activity indicators (e.g. text messages, phone calls) to assess the differences in the weekly patterns of the three cities of London, New York, and Hong Kong. Looking only at the phone call patterns, it shows that the pattern for Hong Kong has a far higher afternoon peak than morning peak, which can also be seen in in our case in Melbourne for user type C. In opposition to that and showing similarities with our case in Sydney for user type C is the phone call pattern for London, which has an equally high morning and afternoon peak. When looking at other patterns such as the weekly text message pattern, different patterns for each of the three cities can be seen.

A third study by Reades et al. (2007) highlights that, besides different topics and cities, also different areas within a city (Rome) show differences in their temporal patterns (Figure 6.1). They showed that for Saturday and Sunday, the activity values often drop far below the typical loadings for the weekdays, whereas the rate of change is also dependent on the area surveyed. They further showed that areas with more residents display lower variances between the individual days than areas with more transient populations such as tourists or commuters. Based on that they suggest that the greater the flux of people is in an area, the greater the variance is in its activity signal.

These findings, however, cannot be linked back to our study, due to the limited hourly number of SSE points in the individual SA2 areas. The aggregated weekly pattern of St Kilda, Melbourne, can serve as an example (Figure 6.2). When looking at both the absolute numbers as well as the pattern itself, it becomes clear that we are presented with a relative small number of SSE points for the individual 2-hour periods. Due to that, a thorough investigation into the differences between different areas and different user types does not make sense, since it cannot be determined whether the differences between those found patterns are randomly generated or not.

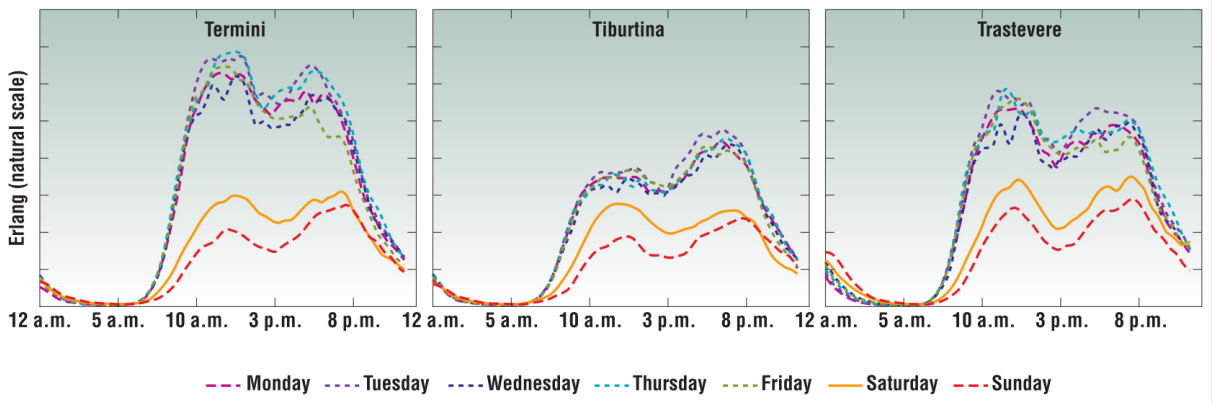

Figure 6.1: Temporal analysis of the mobile phone activity in different areas of Rome (Reades et al. 2007, p.34)

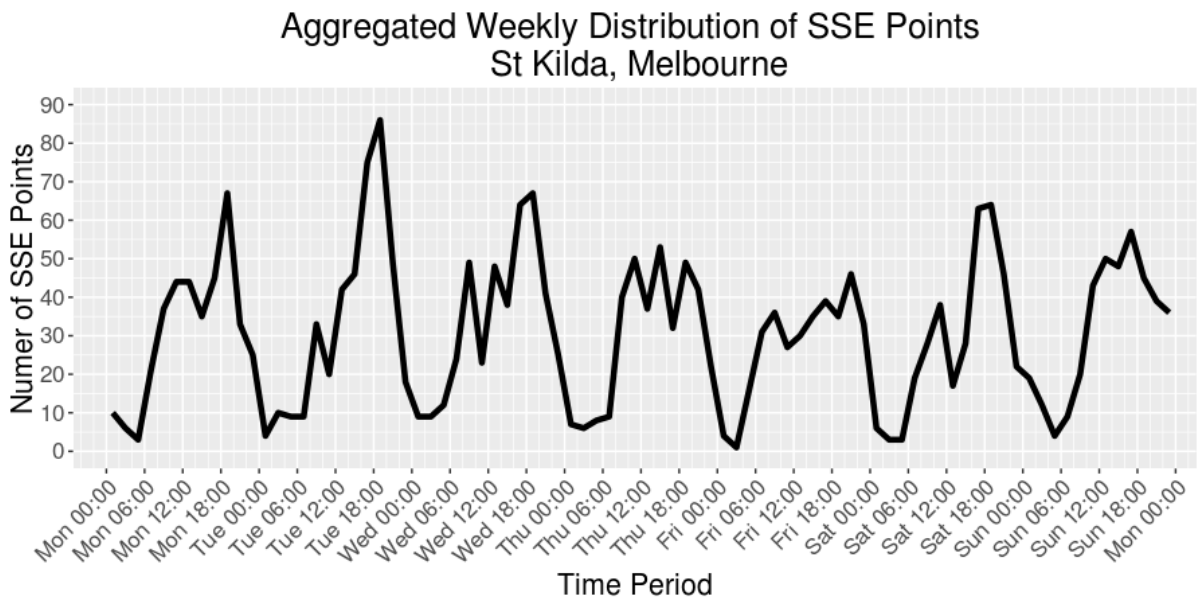

Figure 6.2: Aggregated weekly distribution of SSE points in St Kilda, Melbourne in two-hour windows (absolute numbers).

Periodicity of User Types

In section 5.1.3, we applied Fourier Transformation on the temporal distribution of the SSE points to assess the degree of periodicity for each user type in each city. This showed that although having a more stable daily pattern, the two user types T and E have a smaller periodicity in their aggregated weekly patterns than user type C. Accordingly, user type C then shows higher weekly periodicity and lower daily periodicity.

The identified periodicity reflects the nature of the different user types identified. User type C (Commuter), appears to use the app especially during commuting trips and therefore mostly on weekdays and less on weekends. Accordingly, user type C manifests a small daily periodicity, due to the difference between weekdays and weekends. Furthermore, users of type C use the app more regularly over the course of several weeks which is reflected in the high weekly periodicity.

Users of type T (tourist) are less bound to a working week and therefore do not show significant differences between weekdays and weekends, additionally shown in a higher daily periodicity value. Since they only use the app over a limited amount of time and have a more unstable pattern over several weeks, the weekly periodicity value is rather small. The third user type, type E, manifests patterns between the two extremes of user type C and T. For both the daily as well as the weekly pattern, users of type E manifest periodicity values between the two extremes of user types T and C. The described patterns therefore reflect the assigned labels of the different user types.

Summary and Outlook

Based on these findings we can answer the temporal part of the second research question and conclude this section. We have shown that different user types show different temporal characteristics in both Sydney and Melbourne. The unity of all touristic users (user type T) tends to have a more stable temporal activity pattern when looking at a

weekly aggregation. They also tend to visit the cities with a higher likelihood than other user types, but over a shorter period.

The excursionist users (user type E) show a similar pattern to user type T, whereas the commuter users (user type C) are more active on weekdays than on weekends. Furthermore, we have shown that, although having a smaller daily regularity, user type C has a higher periodicity regarding the weekly patterns. In contrast, user types E and T barely show any weekly periodicity.

A validation of these findings is again rather difficult due to the unknown ground truth of the individual users. A reflection of the found patterns, however, shows that they may be indeed typical for the found user types, especially for user types T and C. Nevertheless, it would be interesting to see whether similar patterns arise from different data sets such as CDR data of the whole city.

We can further state that it is possible to use human navigation data such as the one presented in this thesis to elaborate differences between different users regarding their temporal patterns. This type of data, however, is not useful when trying to elaborate temporal differences between areas and user types, since not enough users are using the app in the same area at the same time. To properly investigate these problems, data with an even higher number of users would be more suited.

6.2.2 Spatial Characteristics

Evaluation of the used Visualization Methods

We used five different methods to assess the spatial characteristics of each user type for each city. Except for the global spatial autocorrelation test, we have principally relied on a visual exploratory analysis. To visualize the various patterns, we used the SA2 areas as described in section 3.4.1. We have further tested other visualizations with SA1-areas (areas are smaller than SA2) and SA3-areas (larger). We have shown that using smaller or larger areas did not allow us to make distinctions between different areas in the same way as it was possible with SA2 areas. It must be clarified, however, that the used areas are based on statistical areas and are not necessarily consistent with the actual human mobility patterns. It therefore underlies the modifiable areal unit problem. Another possible, but not necessarily better approach would have been to use raster cells instead of SA2 areas, similar to the work presented by Reades et al. (2007).

Absolute, Relative Distribution and Location Quotient

We started with an absolute view of the SSE point patterns only depicting the respective number of SSE points/km² per area (section 5.2.1). Although it gives a nice overview of the absolute distribution of the points, this approach has been proven to be rather unsuitable for the comparison of the different user types.

We therefore normalized the percentage of SSE points in each area that belongs to a certain user type (section 5.2.2), since we were looking at the relative amount per area. This approach as well as the location quotient approach allow us to identify areas in which certain user types are over- or underrepresented. We could show that although user type C has the highest percentage of SSE points in all areas, there are certain areas, especial-

ly the highlighted ones, in which user type C is underrepresented. A contrasting image to that is then given by user type T, which is mostly overrepresented in areas in which user type C is underrepresented.

To validate these findings, we assessed the differences between the local percentage values of SSE points belonging to a certain user type and the overall mean value for each area and user type, i.e. the location quotient. The results of this approach (**section 5.2.3**) confirms the findings of the previous section, namely that certain user types are largely under- or overrepresented in certain areas, mostly, again the highlighted ones. The histograms in Figure 5.9 and Figure 5.10 further show that user type C is spatially much more homogeneously distributed in the two cities, whereas the two other user types show much more extreme values, i.e. areas with low or high visiting values. Regarding these two approaches, we must, however, clarify that under-/overrepresentation only refers to the relative amount of SSE points. Accordingly, a certain user type could still have the highest amount of SSE points per area, but is rather underrepresented due to the large difference between the expected mean percentage value and the actually observed percentage value.

Global Spatial Autocorrelation

To test whether the patterns shown in **section 5.2.2** are randomly generated or not, we further tested for global spatial autocorrelation in **section 5.2.4**. The results of the Moran's I test showed that the different patterns (for each city and user type) are not randomly generated, and therefore are statistically significant. This confirmed the findings of the two previous sections, namely that there are certain areas which show differences in the percentage values for different user types and are of a special interest to these user types.

Connectivity

The further realized visualization of the connectivity between SA2 areas for each user type and city gave additional information about user types that had not been known before. Users of type T, for example, not only visit less areas more frequently, these areas are also more connected to each other than to other areas. Furthermore, the areas of user type T that show a high connectivity value are, mostly, again the highlighted areas (city center, airport, beaches). In opposition, users of type C visit many more areas and therefore show much more areas that are connected to each other. User type E then shows a pattern in-between the extreme patterns of user types T and C.

Based on these findings, we can argue that users of type T tend to visit a smaller amount of locations when they visit cities. This reflects also the findings of the temporal analysis, in which we showed that users of type T have a higher probability to visit cities, but spend less time in them than other user types. Accordingly, type T users only visit the areas that are of interest for them: the highlighted areas. On the contrary, type C users that visit or live in the cities spend more time there and therefore visit much more areas. This is then reflected by much more areas that are connected to each other.

Characteristics of the Highlighted Areas

In the various maps depicted in **section 5.2** of the results chapter, we highlighted several areas of each city due to their special patterns found in the maps, especially for user types C and T. In Melbourne, we identified the City Center, the Airport as well as St Kilda. In the case of Sydney, again the City Center and the Airport are highlighted, but also two areas with famous beaches, Bondi Beach and Manly Beach (Destination NSW 2017).

As it can be seen in **section 5.2.5**, commuter type C is relatively underrepresented in the highlighted areas whereas the rather touristic user type T is overrepresented in those areas. Furthermore, the results of the connectivity approach (**section 5.2.5**) shows that for user type T the highlighted areas are more connected to each other than to other areas. Based on these observations, how can these areas be characterized?

Due to their nature (City Center, Beaches, Airport), the highlighted areas can certainly be characterized as areas of certain interest to tourists. This assumption is further strengthened when consulting the comparison of the maps in Figure 6.3. The four maps show that areas with a high density of hotels coincide to a large part with areas in which the touristic user type T is overrepresented (in blue), i.e. has a high location quotient. Tourists are dependent on hotels as a possibility of accommodation and, accordingly, several things can be argued. Firstly, hotels tend to be in areas with a rather large amount of tourist attractions and secondly, tourists tend to visit areas with a high density of tourist attractions. Based on that, we argue that the highlighted areas reflect typical touristic areas.

Especially for Melbourne, our assumptions are strengthened by previous studies. Edwards & Griffin (2013) surveyed the spatial behavior of tourists within Sydney and Melbourne using both interviews and GPS tracking devices. They discovered that certain areas within the two cities are very intensively used by tourists whereas the adjacent areas are mostly not visited at all. They further showed that tourists show repetitive movements. They revisit certain places and use the same routes throughout the day. Furthermore, we have shown that most tourists use the same routes. Although Edwards & Griffin (2013) surveyed tourists on foot and only within the city center, similar patterns can also be seen in the results of the connectivity investigation (**section 5.2.5**). There, we have shown that for user type T only a few areas (Airport, City Center, and Beaches) show strong connections between them, whereas only a few smaller connections between other areas exist.

Another example confirming our assumption is presented by Miah et al. (2016). They used geotagged photos of tourists to assess the popularity of different areas in Melbourne. Using that approach, they identified the Melbourne CBD (City Center) as well as St Kilda and Brighton Beach as the most popular tourist areas. These areas again reflected by high location quotients for user type T.

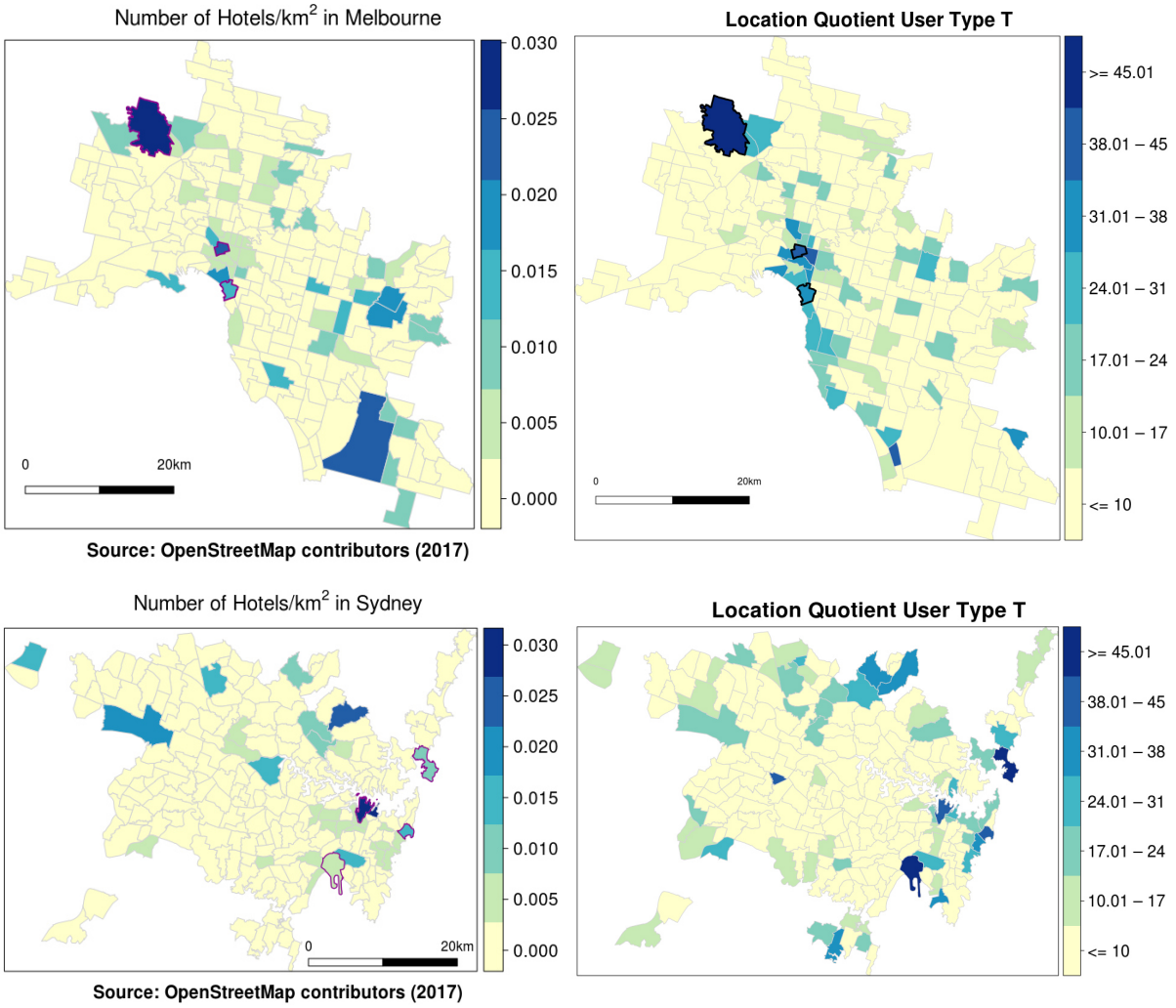

Figure 6.3: Left: Number of Hotels per square kilometer for Melbourne (top) and Sydney (bottom), (OpenStreetMap contributors 2017). Right: Location quotients for each SA2 area, whereas location quotients smaller than 10 are colored with the same color in order to better compare the hotels and location quotients.

Summary and Outlook

Based on the findings of the spatial analysis of the different user types, we can answer the second research question and conclude this section. We have shown that different user types overall visit similar places, although with different likelihoods and magnitudes. We have further shown that the visiting pattern of certain areas are, relatively seen, dominated by certain user types. In areas that can be interpreted as tendential touristic areas, more SSE points of users belonging to the touristic user type T can be found. Furthermore, less areas are visited by the touristic users, whereas users belonging to the commuter type C visit many more areas, shown by the higher connectivity of the different areas. We were further able to show that areas with a relative overrepresentation of SSE points belonging to user type T are areas that show a high concentration of hotels.

We are aware however, that a characterization of the individual areas solely based on the found patterns, can only generate very vague results. For a more detailed analysis and characterization of the individual areas, more data from various data sources as well as a qualitative assessment of domain experts would be needed.

Accordingly, validation of the found patterns for each user type is complex due to the explorative approach of this thesis. January is not only a popular month with tourists coming from abroad, but also for Australians themselves to go on holidays. Due to that, we have applied various methods to find specific patterns for each user type. Nevertheless, we must point out that the pattern found may present itself differently when looking at other months or several other months combined.

7. Conclusion

7.1 Summary

There have been many approaches to discover patterns in human movement behavior based on positioning data with a high temporal and spatial resolution. GPS navigation data from a mobile phone application as used in this study, however, are not among the used data sources. This kind of data presents both high spatial and temporal resolution and is unique do to the reason people use the app; navigation and the need of additional (spatial) information. Accordingly, we argue that a lot of the people using the app are unfamiliar with the surrounding and locations they visit. We therefore believe that a high proportion of users are tourists of some sort.

Besides the novel data source, also the applied analysis of human movement in cities differs much from previous studies. Mostly, users have been handled as a similar-acting and homogeneous community, due to several reasons. Either, the surveyed community itself was very homogenous, or, only communities were survey for which additional information was available. The here used navigation data, however, does not offer that. We have therefore used that handicap to our advantage and presented a novel methodology to divide users into similar acting user types, based on their spatio-temporal footprints and in the absence of ground truth. Accordingly, we neglect findings of previous studies and use the differences of the users in our study to gain a deeper understanding of the individual types of users that use a navigation app.

We have shown that the identified user types show distinct temporal and spatial patterns found in their start, stop and end points of trajectories. The examined patterns in Melbourne and Sydney moreover reflect the nature of the found user types. On the one hand, commuter users tend to cover larger areas within the city itself and show a wide spreading pattern. On the other hand, users showing touristic movement patterns more often visit only small numbers of different areas, whereas these areas reflect touristic hotspots such as city centers, airports, or beaches.

7.2 Contributions

This thesis presents a contribution to human mobility research that is comparatively untested, by characterizing users based on their spatio-temporal movement behavior in the absence of ground truth. Two unsupervised machine learning techniques, clustering and PCA, have been applied on various measures describing the individuals' spatio-temporal footprints in order to find most distinct groups, i.e. user types. For the clustering, four different methods with several numbers of principle components have been tested and validated to find the most suitable combination of clustering method and clustering parameters, i.e. the combination of the number of principal components to use and the clustering method itself. A qualitative analysis of the clusters found was then used to discover and interpret distinct user types. By analyzing the temporal and spatial characteristics of these user types, we have further shown that the elaborated user types disproportional visit certain areas that often reflect the nature of the individual user types.

Identifying such different user types in the absence of ground truth can be of assistance for various other research areas. The identification of users with touristic behaviors for example can be directly used by Sygic to provide better navigation help for people unfamiliar with their surroundings. On the other hand, the identification of the commuter users can be used to provide them with specially tailored navigation products. Besides the navigation domain, the identification of touristic users also helps in the investigation of virus spreading patterns, since touristic users tend to visit a bigger variety of locations and areas, and therefore are prone to spread viruses over bigger areas. Other than that, the findings of our applied methodology can also be of use for new applications in the tourist management and urban planning domain. Since the high spatial and temporal resolution of the data offers new insights into the movement behavior of various user types, they can be used to detect areas with a high potential for certain services.

We further see a potential of the presented methodology in insurance and healthcare areas. Certain user types might be more prone to accidents due to their unfamiliarity with the visited areas and show a more precarious driving behavior. Identifying such user types would help insurance companies to tailor the products more efficiently based on the driving behavior.

7.3 Outlook

In order to further validate the presented methodology, a testing on a larger and different data set over several months and over individual months compared is required. By applying the methodology on several months, we are presented with an even better movement behavior of the different users. When further applying the presented methodology on individual months separately, we would be able to establish even more detailed user types. Accordingly, we could test whether we could learn a model based on data for one month and then try to predict the class labels of the individual users for the next or a different month. A further approach would be to generate a model that would instantly label users based on their just recorded movements.

Furthermore, it would be interesting to see whether the methodology, i.e. the trained method can be directly applied in a different geographic context, i.e. different countries individually. It would further be interesting to have a look at data from areas with a lot of border crossings, since we are not presented with such cases due to the isolation of Australia. Applying such a methodology on data of these areas would additionally help to investigate whether users of different background, i.e. nationalities show different spatio-temporal footprints and behaviors.

An area for further examination could be a sensitivity analysis of different approaches to segment the trajectories. In this thesis, we have used only one spatial and temporal threshold to detect stops. It could therefore be interesting to see whether the found patterns differ when using different thresholds.

An additional interesting approach would be to cross-reference the patterns found with other data sets, such as CDR data or RFID data of public transport usage in the two cities. We could then check whether the found patterns also occur in these data sets or whether they differ, and if so, how they differ and why. We would further be interested to analyze whether the different user types follow the paths given by the navigation app or whether they refuse the app's suggestions and take their own paths.

References

Ahas, R. et al., 2008. Evaluating Passive Mobile Positioning Data for Tourism Surveys: An Estonian Case Study. *Tourism Management*, 29(3), pp.469–486.

Ahas, R. et al., 2015. Everyday Space–Time Geographies: Using Mobile Phone-Based Sensor Data to Monitor Urban Activity in Harbin, Paris, and Tallinn. *International Journal of Geographical Information Science*, 29(11), pp.2017–2039.

Ahas, R. et al., 2014. *Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics - Consolidated Report* Positium LBS, ed., Eurostat.

Ahas, R. et al., 2007. Seasonal Tourism Spaces in Estonia: Case Study with Mobile Positioning Data. *Tourism Management*, 28(3), pp.898–910.

Alvares, L.O. et al., 2007. A Model for Enriching Trajectories with Semantic Geographical Information. In *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems (GIS '07)*. pp. 162–169.

Andrienko, G. et al., 2013. Extracting Semantics of Individual Places from Movement Data by Analyzing Temporal Patterns of Visits. In *Proceedings of The First ACM SIGSPATIAL International Workshop on Computational Models of Place - COMP '13*. New York, New York, USA: ACM Press, pp. 9–16.

Andrienko, N., Andrienko, G. & Rinzivillo, S., 2015. Exploiting Spatial Abstraction in Predictive Analytics of Vehicle Traffic. *ISPRS International Journal of Geo-Information*, 4(2), pp.591–606.

Australian Bureau of Statistics, 2016a. Australian Statistical Geography Standard (ASGS). *Statistical Geography*. Available at: http://www.abs.gov.au/websitedbs/d3310114.nsf/home/australian+statistical+geography+standard+(asgs) [Accessed March 7, 2017].

Australian Bureau of Statistics, 2015. Estimated Resident Population - Greater Capital City Statistical Areas (GCCSAs). *3218.0 - Regional Population Growth, Australia, 2014-15*. Available at: http://www.abs.gov.au/ausstats/abs@.nsf/0/797F86DBD192B8F8CA2568A9001393CD?Opendocument [Accessed February 27, 2017].

Australian Bureau of Statistics, 2016b. Overseas Arrivals and Departures, Australia, May 2016. *3401.0 - Overseas Arrivals and Departures, Australia, May 2016*. Available at: http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/3401.0Main+Features1May2016?OpenDocument [Accessed November 14, 2016].

van den Berg, R.A. et al., 2006. Centering, Scaling, and Transformations: Improving the Biological Information Content of Metabolomics Data. *BMC Genomics*, 7(142), pp.1–15.

Biljecki, F., Ledoux, H. & van Oosterom, P., 2013. Transportation Mode-Based Segmentation and Classification of Movement Trajectories. *International Journal of Geographical Information Science*, 27(2), pp.385–407.

Birenboim, A. & Shoval, N., 2015. Mobility Research in the Age of the Smartphone. *Annals of the Association of American Geographers*, 106(2), pp.283–291.

Bivand, R., Pebesma, E. & Gomez-Rubio, V., 2013. *Applied Spatial Data Analysis with R* 2. Edition., New York: Springer-Verlag New York.

Bivand, R. & Piras, G., 2015. Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software*, 63(18), pp.1–36.

Bivand, R. & Rundel, C., 2016. rgeos: Interface to Geometry Engine - Open Source (GEOS). R package version 0.3-21.

Bro, R. & Smilde, A.K., 2003. Centering and Scaling in Component Analysis. In *Journal of Chemometrics*. pp. 16–33.

Bro, R. & Smilde, A.K., 2014. Principal Component Analysis. *Analytical Methods*, 6(9), pp.2812–2831.

Brock, G. et al., 2008. clValid: An R Package for Cluster Validation. *Journal Of Statistical Software*, 25(March 2008), pp.1–28.

Brunsdon, C. & Comber, L., 2015. *An Introduction to R for Spatial Analysis and Mapping* 1. Edition. R. Rojek et al., eds., London: SAGE Publications Ltd.

Calabrese, F. et al., 2013. Understanding Individual Mobility Patterns from Urban Sensing Data: A Mobile Phone Trace Example. *Transportation Research Part C: Emerging Technologies*, 26, pp.301–313.

Calabrese, F., Reades, J. & Ratti, C., 2010. Eigenplaces: Segmenting Space Through Digital Signatures. *IEEE Pervasive Computing*, 9(1), pp.78–84.

Candia, J. et al., 2008. Uncovering Individual and Collective Human Dynamics from Mobile Phone Records. *Journal of Physics A: Mathematical and Theoretical*, 41(22), pp.1–11.

Cheng, J. & Xie, Y., 2016. leaflet: Create Interactive Web Maps with the JavaScript "Leaflet" Library. R package version 1.0.1.

Conway, J. et al., 2016. RPostgreSQL: R interface to the PostgreSQL database system. R package version 0.4-1.

Das, R.D., Ronald, N. & Winter, S., 2015. A Simulation Study on Automated Transport Mode Detection in Near-Real Time Using a Neural Network. In *CEUR Workshop Proceedings*. CEUR-WS, pp. 46–57.

Datta, S. & Datta, S., 2003. Comparisons and Validation of Statistical Clustering Techniques for Microarray Gene Expression Data. *Bioinformatics*, 19(4), pp.459–466.

Demissie, M.G., de Almeida Correia, G.H. & Bento, C., 2013. Intelligent Road Traffic Status Detection System through Cellular Networks Handover Information: An Exploratory Study. *Transportation Research Part C: Emerging Technologies*, 32, pp.76–88.

Department of Infrastructure and Transport, 2012. *Traffic Growth in Australia*, Canberra ACT.

Destination NSW, 2017. Sydney Beaches. *Sydney*. Available at: http://www.sydney.com/things-to-do/beach-lifestyle/sydney-beaches [Accessed April 11, 2017].

Dodge, S. et al., 2016. Analysis of Movement Data. *International Journal of Geographical Information Science*, 30(5), pp.825–834.

Dodge, S., Laube, P. & Weibel, R., 2012. Movement Similarity Assessment using Symbolic Representation of Trajectories. *International Journal of Geographical Information Science*, 26(9), pp.1563–1588.

Edwards, D. & Griffin, T., 2013. Understanding Tourists' Spatial Behaviour: GPS Tracking as an Aid to Sustainable Destination Management. *Journal of Sustainable Tourism*, 21(4), pp.580–595.

Esri, 2016. How Spatial Autocorrelation (Global Moran's I) works. *ArcMap 10.3*. Available at: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/h-how-spatial-autocorrelation-moran-s-i-spatial-st.htm [Accessed April 6, 2017].

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P., 1996. *Advances in Knowledge Discovery and Data Mining*, AAAI Press.

Fennell, D.A., 1996. A Tourist Space-Time Budget in the Shetland Islands. *Annals of Tourism Research*, 23(4), pp.811–829.

Filion, P., 2000. Balancing Concentration and Dispersion? Public Policy and Urban Structure in Toronto. *Environment and Planning C: Government and Policy*, 18(2), pp.163–189.

Giannotti, F. et al., 2011. Unveiling the Complexity of Human Mobility by Querying and Mining Massive Trajectory Data. *The VLDB Journal*, 20(5), pp.695–719.

Girardin, F. et al., 2008. Digital Footprinting: Uncovering Tourists with User-Generated Content. *IEEE Pervasive Computing*, 7(4), pp.36–43.

Girardin, F. et al., 2009. Quantifying Urban Attractiveness from the Distribution and Density of Digital Footprints. *International Journal of Spatial Data Infrastructures Research*, 4(4), pp.175–200.

González, M.C., Hidalgo, C.A. & Barabási, A.-L., 2008. Understanding Individual Human Mobility Patterns. *Nature*, 453(7196), pp.779–82.

Grauwin, S. et al., 2015. Towards a Comparative Science of Cities: Using Mobile Traffic Records in New York, London, and Hong Kong. In *Computational Approaches for Urban Environments*. Springer International Publishing, pp. 363–387.

Grolemund, G. & Wickham, H., 2011. Dates and Times Made Easy with lubridate. *Journal of Statistical Software*, 40(3), pp.1–25.

Halkidi, M., Batistakis, Y. & Vazirgiannis, M., 2001. On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17(2–3), pp.107–145.

Han, J., Kamber, M. & Pei, J., 2012. *Data Mining: Concepts and Techniques* 3. Edit., Elsevier Inc.

Hasan, S. et al., 2013. Spatiotemporal Patterns of Urban Human Mobility. *Journal of Statistical Physics*, 151(1–2), pp.304–318.

Hastie, T., Tibshirani, R. & Friedman, J.H., 2009. *The Elements of Statistical Learning* 2. Edition., New York, NY: Springer Verlag.

Jain, A.K., 2010. Data Clustering: 50 Years beyond K-Means. *Pattern Recognition Letters*, 31(8), pp.651–666.

James, G. et al., 2013. *An Introduction to Statistical Learning*, New York, New York, USA: Springer-Verlag New York.

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, pp.122–135.

Jiang, S. et al., 2013. A review of Urban Computing for Mobile Phone Traces. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13*. New York, New York, USA: ACM Press, pp. 1–9.

Jiang, S., Ferreira, J. & González, M.C., 2015. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. *IEEE Transactions on Big Data*, pp.1–13.

Kassambara, A. & Mundt, F., 2016. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.3.

Kling, F. & Pozdnoukhov, A., 2012. When a City Tells a Story. In *Proceedings of the 20th International Conference on Advances in Geographic Information Systems - SIGSPATIAL '12*. New York, New York, USA: ACM Press, pp. 482–485.

Kvalheim, O.M., Brakstad, F. & Liang, Y., 1994. Preprocessing of Analytical Profiles in the Presence of Homoscedastic or Heteroscedastic Noise. *Analytical Chemistry*, 66, pp.43–51.

Laube, P., 2014. *Computational Movement Analysis* 1. Edition., Wädenswil: Springer International Publishing.

Lee, J. et al., 2008. TraClass: Trajectory Classification Using Hierarchical Region Based and Trajectory Based Clustering. *Proceedings of the VLDB Endowment*, 1(1), pp.1081–1094.

Leng, Y. et al., 2016. Analysis of Tourism Dynamics and Special Events through Mobile Phone Metadata. In *Proceedings of Data for Good Exchange (D4GX) 2016*. New York, NY, p. 6.

Maechler, M. et al., 2015. Cluster: cluster analysis basics and extensions. R package version 2.0.5.

Maps of World, 2013. Australia Map. *Australia*. Available at: http://www.mapsofworld.com/australia/ [Accessed March 29, 2017].

Miah, S.J. et al., 2016. A Big Data Analytics Method for Tourist Behaviour Analysis. *Information & Management*.

de Montjoye, Y.-A. et al., 2013. Unique in the Crowd: The Privacy Bounds of Human Mobility. *Scientific reports*, 3(1376), pp.1–5.

Neuwirth, E., 2014. RColorBrewer: ColorBrewer Palettes. R package version 1.1-2.

Noulas, A. et al., 2012. A Tale of Many Cities: Universal Patterns in Human Urban Mobility. *PloS one*, 7(5), pp.1–10.

OpenStreetMap contributors, 2016. Elements. *OpenStreetMap Wiki*. Available at: https://wiki.openstreetmap.org/wiki/Elements.

Palchykov, V. et al., 2014. Inferring Human Mobility using Communication Patterns. *Scientific Reports*, 4(6174), pp.1–6.

Panda, S., Hogenesch, J.B. & Kay, S.A., 2002. Circadian Rhythms from Flies to Human. *Nature*, 417(6886), pp.329–335.

Pappalardo, L. et al., 2013. Understanding the Patterns of Car Travel. *The European Physical Journal Special Topics*, 215(1), pp.61–73.

Parent, C. et al., 2013. Semantic Trajectories Modeling and Analysis. *ACM Computing Surveys*, 45(4), pp.1–32.

Pei, T. et al., 2014. A new Insight into Land Use Classification based on Aggregated Mobile Phone Data. *International Journal of Geographical Information Science*, 28(9), pp.1988–2007.

Pelekis, N. et al., 2012. Visually Exploring Movement Data via Similarity-Based Analysis. *Journal of Intelligent Information Systems*, 38(2), pp.343–391.

Phithakkitnukoon, S. et al., 2010. Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data. In *Lecture Notes in Computer Science*. pp. 14–25.

Pihur, V., Datta, S. & Datta, S., 2009. RankAggreg: an R Package for Weighted Rank Aggregation. R package version 0.5. *BMC Bioinformatics*, 10(62), pp.1–10.

PostGIS, 2016. ST_ConcaveHull. *PostGIS Docs*. Available at: http://postgis.net/docs/ST_ConcaveHull.html [Accessed February 23, 2017].

R Core Team, 2016. R: A Language and Environment for Statistical Computing.

Ranacher, P. et al., 2016. What is an Appropriate Temporal Sampling Rate to Record Floating Car Data with a GPS? *ISPRS International Journal of Geo-Information*, 5(1), pp.1–17.

Reades, J. et al., 2007. Cellular Census: Explorations in Urban Data Collection. *IEEE Pervasive Computing*, 6(3), pp.30–38.

Reades, J., Calabrese, F. & Ratti, C., 2009. Eigenplaces: Analysing Cities using the Space - Time Structure of the Mobile Phone Network. *Environment and Planning B: Planning and Design*, 36(5), pp.824–836.

Ren, Y. et al., 2016. D-Log: A WiFi Log-Based Differential Scheme for Enhanced Indoor Localization with Single RSSI Source and Infrequent Sampling Rate. *Pervasive and Mobile Computing*.

Renso, C. et al., 2012. How You Move Reveals Who You Are: Understanding Human Behavior by Analyzing Trajectory Data. *Knowledge and Information Systems*, 37(2), pp.331–362.

Rousseeuw, P., 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, pp.53–65.

Sagl, G., Loidl, M. & Beinat, E., 2012. A Visual Analytics Approach for Extracting Spatio-Temporal Urban Mobility Information from Mobile Network Traffic. *ISPRS International Journal of Geo-Information*, 1(3), pp.256–271.

Shoval, N. et al., 2011. Hotel Location and Tourist Activity in Cities. *Annals of Tourism Research*, 38(4), pp.1594–1612.

Shoval, N. & Ahas, R., 2016. The Use of Tracking Technologies in Tourism Research: A Review of the First Decade. *Tourism Geographies*, 18(5), pp.1–20.

Shoval, N. & Isaacson, M., 2007. Tracking Tourists in the Digital Age. *Annals of Tourism Research*, 34(1), pp.141–159.

Silm, S. & Ahas, R., 2010. The Seasonal Variability of Population in Estonian Municipalities. *Environment and Planning A*, 42(10), pp.2527–2546.

Spaccapietra, S. et al., 2008. A Conceptual View on Trajectories. *Data and Knowledge Engineering*, 65(1), pp.126–146.

Spaccapietra, S. & Parent, C., 2011. Adding Meaning to Your Steps. In *Lecture Notes in Computer Science*. pp. 13–31.

Stenneth, L. et al., 2011. Transportation Mode Detection Using Mobile Phones and GIS Information. In *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '11*. New York, New York, USA: ACM Press, p. 54.

Sygic, 2016. About Sygic. Available at: http://www.sygic.com/about [Accessed November 14, 2016].

Tan, P.-N., Steinbach, M. & Kumar, V., 2006. *Introduction to Data Mining*, Pearson Addison Wesley.

The World Bank, 2015. Urban population (% of total). *World Development Indicators*. Available at: http://data.worldbank.org/indicator/SP.URB.TOTL.IN.ZS?order=wbapi_data_value_2014+wbapi_data_value+wbapi_data_value-last&sort=desc [Accessed May 23, 2016].

Thomason, A., Griffiths, N. & Sanchez, V., 2016. Context Trees: Augmenting Geospatial Trajectories with Context. *ACM Transactions on Information Systems*, 35(2), pp.1–37.

Tibshirani, R., Walther, G. & Hastie, T., 2001. Estimating the Number of Clusters in a Data Set via the Gap Statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63, pp.411–423.

Tietbohl, A. et al., 2008. A Clustering-Based Approach for Discovering Interesting Places in Trajectories. In *23rd Annual ACM Symposium on Applied Computing, SAC'08*. pp. 863–868.

Toole, J.L. et al., 2012. Inferring Land Use from Mobile Phone Activity. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. pp. 1–8.

Tourism Australia, 2017. Drive Australia. *Itineraries*. Available at: http://www.australia.com/en/itineraries/drive-australia.html [Accessed February 20, 2017].

Tourism Research Australia, 2016. *National Visitor Survey Results*, Canberra ACT.

Trasarti, R. et al., 2011. Mining Mobility User Profiles for Car Pooling. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*. New York, New York, USA: ACM Press, pp. 1190–1198.

Ubicomp@UMinho, 2006. Concave Hull. *LOCAL - Location Contexts for Location-Aware Applications*. Available at: http://ubicomp.algoritmi.uminho.pt/local/concavehull.html [Accessed February 23, 2017].

Vajakas, T., Vajakas, J. & Lillemets, R., 2015. Trajectory Reconstruction from Mobile Positioning Data using Cell-to-Cell Travel Time Information. *International Journal of Geographical Information Science*, 8816(June), pp.1–14.

Visit Victoria, 2016. St Kilda. *Destinations*. Available at: http://www.visitvictoria.com/Regions/Melbourne/Destinations/St-Kilda [Accessed April 30, 2017].

Wickham, H., 2009. *ggplot2: Elegant Graphics for Data Analysis*, New York: Springer-Verlag New York.

Wickham, H., 2016. tidyr: Easily Tidy Data with `spread()` and `gather()` Functions. R package version 0.6.0.

Wickham, H. & Francois, R., 2016. dplyr: A Grammar of Data Manipulation. R package version 0.5.0.

Witten, I.H., Frank, E. & Hall, M.A., 2011. *Data Mining: Practical Machine Learning Tools and Techniques* 3. Edition., Morgan Kaufmann Publishers Inc.

Wu, J. et al., 2011. Automated Time Activity Classification based on Global Positioning System (GPS) Tracking Data. *Environmental Health: A Global Access Science Source*, 10(101), pp.1–13.

Yuan, J., Zheng, Y. & Xie, X., 2012. Discovering Regions of Different Functions in a City using Human Mobility and POIs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2012*. New York, New York, USA: ACM Press, pp. 186–194.

Yuan, Y. & Raubal, M., 2012. Extracting Dynamic Urban Mobility Patterns from Mobile Phone Data. *Geographic information science*, 7478 LNCS, pp.354–367.

Zhang, D. et al., 2014. Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales. In *Proceedings of the 20th ACM Annual International Conference on Mobile Computing and Networking - MobiCom 2014*. New York, New York, USA: ACM Press, pp. 201–212.

Zhao, Z. et al., 2016. Understanding the Bias of Call Detail Records in Human Mobility Research. *International Journal of Geographical Information Science*, 30(9), pp.1738–1762.

Zheng, K. et al., 2013. On Discovery of Gathering Patterns from Trajectories. In *Proceedings of the 29th IEEE International Conference on Data Engineering*. Brisbane, QLD, pp. 242–253.

Zheng, Y. et al., 2008. Learning Transportation Mode from Raw GPS Data for Geographic Applications on the Web. In *Proceeding of the 17th International Conference on World Wide Web - WWW '08*. New York, New York, USA: ACM Press, p. 247.

Zheng, Y. et al., 2011. Recommending Friends and Locations based on Individual Location History. *ACM Transactions on the Web*, 5(1), pp.1–44.

Zheng, Y., 2015. Trajectory Data Mining: An Overview. *ACM Transactions on Intelligent Systems and Technology*, 6(3), pp.1–41.

Zheng, Y. et al., 2014. Urban Computing: Concepts, Methodologies, and Applications. *ACM Transactions on Intelligent Systems and Technology*, 5(3), pp.1–55.

Zhou, Z.-H., 2003. Three Perspectives of Data Mining. *Artificial Intelligence*, 143(1), pp.139–146.

Zook, M., Kraak, M.-J. & Ahas, R., 2015. Geographies of Mobility: Applications of Location-Based Data. *International Journal of Geographical Information Science*, pp.1935–1940.

Appendix

A. SA2 Areas Melbourne

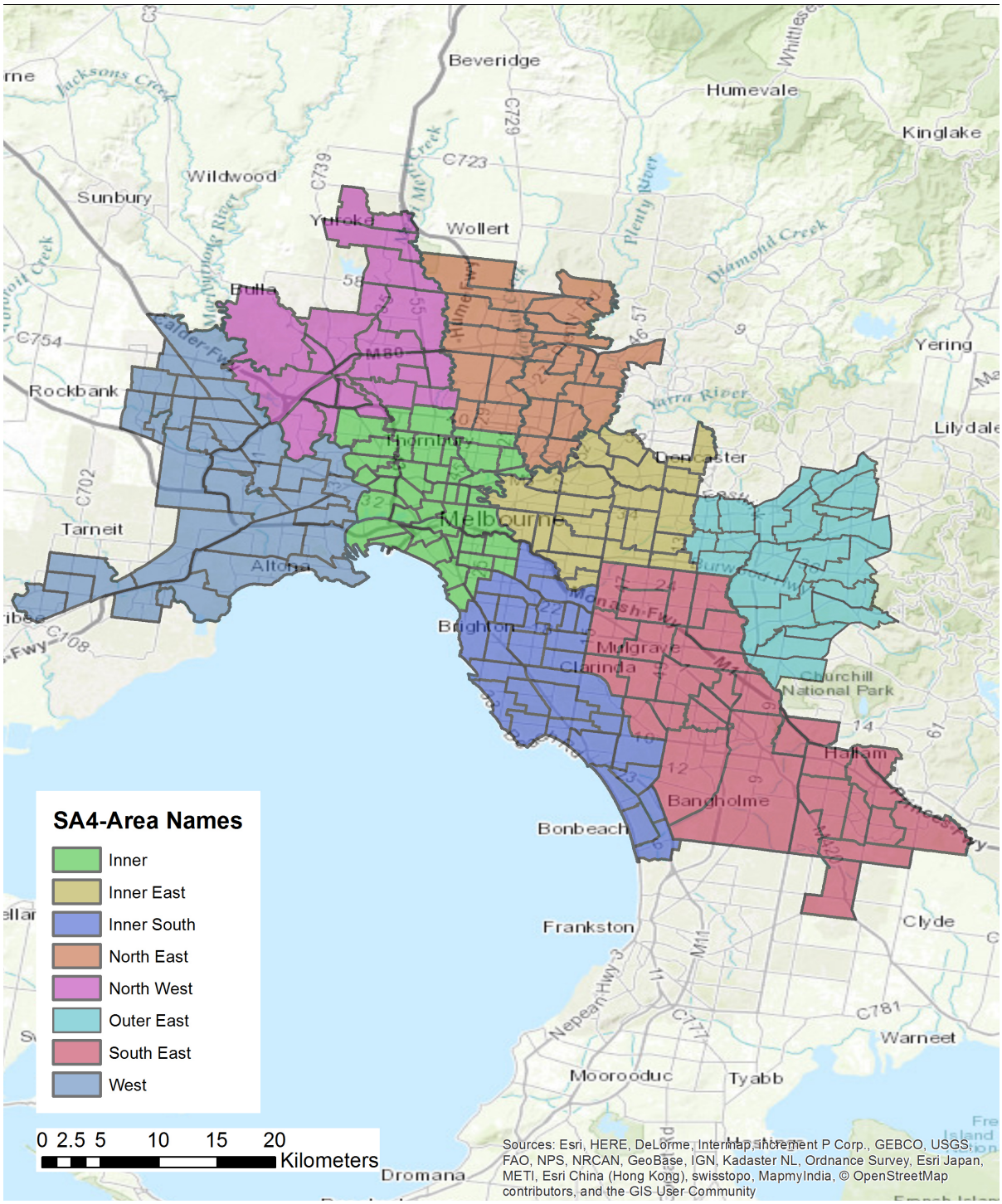

Figure A.1: SA2 areas of Melbourne colored corresponding to their SA4-area name

B. SA2 Areas Sydney

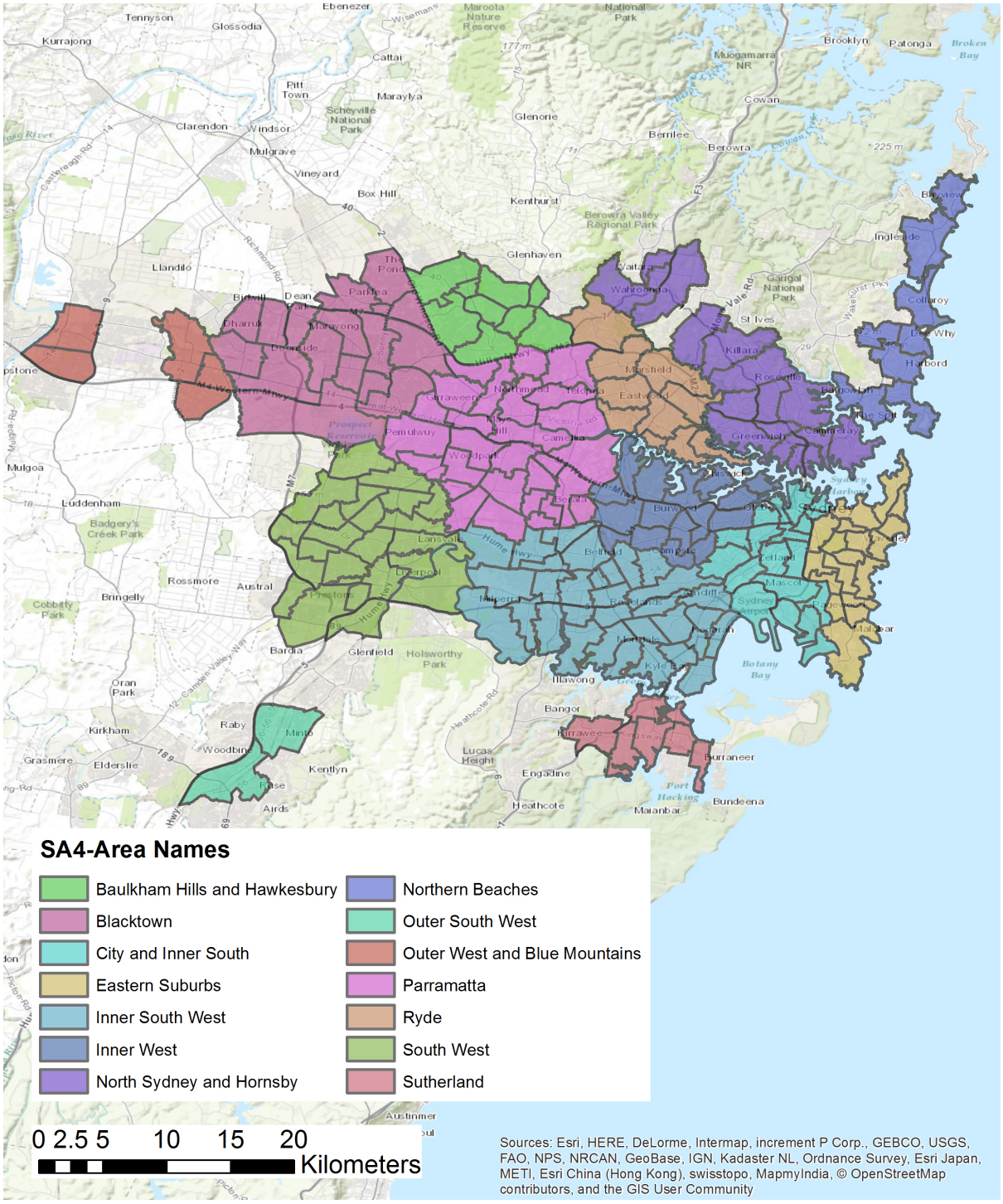

Figure A.2: SA2 areas of Sydney colored corresponding to their SA4-area name

C. Important SQL-Queries

```
CREATE TABLE movingwindow_5m AS
SELECT a.*,
               SUM(b.distance) AS sum_dist_m5,
               count(*) AS count
       FROM points AS a
       JOIN points AS b
               USING (p_sessionid)
       WHERE b.timestamp >= a.timestamp
               AND   b.timestamp < a.timestamp + interval '5 minute'
       GROUP BY a.pointid
       ORDER BY a.timestamp, a.pointid;
```

Code Fragment 1: SQL-Code to calculate the covered distance over a moving window of five minutes.

```
CREATE TABLE stops AS
       SELECT d.*,
               CASE
                       WHEN d.sum_dist_m5 <= 10
                       AND d.count >= 5
                       THEN 1
                       ELSE 0
                       END AS point_type
       FROM movingwindow_5m AS d;
```

Code Fragment 2: SQL-Code to tag the tuples that match the requirements of a stop point.

```
UPDATE sse_point AS k
       SET    osm_type = d.osm_type,
               dist2osm = d.dist2osm
       FROM
               (SELECTa.pointid,
                       b.name AS osm_name,
                       b.type AS osm_type,
                       ST_Distance(a.coord::geography, b.geom::geography) as dist2osm
               FROM
                       (SELECTpointid,
                               coord
                       FROM sse_point) AS a,
                               osm_poi AS b
               WHERE ST_Intersects(a.coord, b.buffer)
               ORDER BY pointid, dist2osm)
               AS d
       WHERE k.pointid = d.pointid;
```

Code Fragment 3: SQL-Code to update the table of the SSE points with information about the nearest OSM POI

```
SELECT d.deviceid,
       a.concave_hull,
       ST_Centroid(a.concave_hull) AS d_centroid         --centroid of concave hull
FROM dailymovement d
LEFT JOIN
     (WITH hulls AS (              --from a table where we first have to check whether the
                                   geometry type is a polygon
              SELECT p_dmoveid,
                     st_concavehull(
                           ST_Collect(coordinates), 0.90) AS concave_hull     --target  per-
                                                                              cent: 90%
              FROM points
              GROUP BY dailymovementid)
       SELECT * FROM hulls where ST_GeometryType(concave_hull)='ST_Polygon') AS a;
```

Code Fragment 4: SQL-Code to compute the daily concave hull and its centroid

```
CREATE TABLE sse_clusters AS
       SELECT row_number() over () AS cluster_id,        --a cluster id
              deviceid,
              ST_NumGeometries(cluster) AS number_points,       --no. of points in cluster
              ST_Centroid(clust) AS centroid,                   --centroid of the cluster
              ST_MinimumBoundingCircle(cluster) AS circle,      --minimum bounding circle
              sqrt(ST_Area(ST_MinimumBoundingCircle(cluster))
                     / pi()) AS radius            --radius of minimum bounding circle
       FROM
              (SELECT
                     unnest(
                            ST_ClusterWithin(                          --calculate cluster
                                   coordinates, 15)) AS cluster,       --eps=15m
                     deviceid
              FROM start_stop_end
              GROUP BY deviceid) AS f;
```

Code Fragment 5: SQL-Code to compute the spatial cluster per user

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.

April 13, 2017

Luca Scherrer