University of Zurich[UZH]

Do Taxi Drivers Take the Shortest Routes?

A Large-Scale Analysis Using GPS-based Floating Car Data in Vienna

GEO 511 – MSc thesis

Author

Darryl Schumacher

12-712-022

Supervisors

Dr. Haosheng Huang

Anita Graser
(AIT Austrian Institute of Technology GmbH
Giefinggasse 42, 1210 Vienna, Austria
E-Mail: anita.graser@ait.ac.at)

Faculty supervisor

Prof. Dr. Robert Weibel

Date of submission: 30.06.2018

Department of Geography, University of Zurich

Contact

Darryl Schumacher
Grundstrasse 7
CH-8180 Bülach, Switzerland
darryl@bluewin.ch

Prof. Dr. Robert Weibel
Geographic Information Science (GIS)
Department of Geography
University of Zurich - Irchel
Winterthurerstr. 190
CH-8057 Zurich, Switzerland
robert.weibel@geo.uzh.ch

Dr. Haosheng Huang
Geographic Information Science (GIS)
Department of Geography
University of Zurich - Irchel
Winterthurerstr. 190
CH-8057 Zurich, Switzerland
haosheng.huang@geo.uzh.ch

Anita Graser
AIT Austrian Institute of Technology GmbH
Center for Mobility Systems
Dynamic Transportation Systems
Giefinggasse 42, 1210 Vienna, Austria
anita.graser@ait.ac.at

Abstract

Understanding the nature of taxi drivers' route choice behavior is essential for traffic modeling as well as the development of intelligent transportation systems. On the other hand, studying how taxi drivers make route decisions will also provide important insights to improve existing car navigation systems, which so far mostly provide shortest or fastest routes. This thesis aims to analyze how taxi drivers make route choice decisions when they have passengers onboard, using a large-scale FCD (floating car data) dataset in Vienna (Austria). Particularly, the following research questions will be addressed: Do taxi drivers with passengers onboard take the fastest or shortest routes? How do the chosen routes differ from fastest and shortest routes? What are the route characteristics preferred by taxi drivers with passengers onboard? How do certain external factors affect the route choice behavior? How do they choose their routes in certain scenarios? What is the influence of network centrality and geography?

The analysis consists of three parts. First, shortest and fastest paths have been calculated for all observed paths. By calculating eight route characteristics, these three groups were then compared with each other. Second, the influence of three external factors – time of the day, day of the week and weather conditions – on observed trips' route characteristics has been analyzed. Last, the influence of network centrality measures – betweenness, closeness and information – and the effect of geography on route choice behavior has been investigated using detailed case studies.

The results show that shortest, fastest and observed paths differ significantly from each other. Furthermore, taxi drivers often do not follow either of the computed paths. The influence of time of the day and day of the week has a clear impact on chosen routes' characteristics. Most importantly, the more traffic there is, drivers use relatively more tertiary roads than during periods with less traffic and the overlap with both groups of computed paths becomes higher. The case studies show that taxi drivers in Vienna do not exhibit a directional bias as they usually choose the same routes going from point A to point B as from B to A. Additionally, the location of the chosen routes differs depending on time of the day. Furthermore, centrality influences route choice behavior. Especially higher betweenness centrality correlates to some extent with the popularity of a road segment. However, none of the three centralities alone can explain the behavior of taxi drivers adequately. Last, route characteristics differ significantly between different areas of the city, indicating that geography and street network configuration have clear influence on chosen routes.

The results suggest that the route choice behavior of taxi drivers cannot be adequately represented with the shortest path model nor the anchor-based model. However, taxi drivers seem to choose routes that are efficient in terms of travel time. Thus, the analysis of taxi drivers' route choice behavior could help to enhance navigation systems and better transportation systems could be developed.

Keywords: route choice, shortest path, trajectories, cluster analysis, floating car data, centrality

Acknowledgments

After a year of very interesting research on route choice behavior, I would like to thank Anita Graser and Haosheng Huang for their valuable input and for their suggestions throughout the duration of the work.

Additionally, I would also like to thank Taxi 31300 for providing me with the floating car data, which made this thesis possible in the first place.

Content

Figures

Tables

ABBREVIATONS

DBMS – Database management system

FCD – Floating car data

GIS – Geographical Information Systems

GIScience – Geographical Information Science

LISA – Local indicator of spatial association

RDI – Route directness index

PSL – Percentage of shared length

OD – Origin and destination

OSM – OpenStreetMap

1. Introduction

Analysis of route choice behavior is a broad research field, that has attracted a lot of interest in recent years for several reasons. First, there is a demand for better and more detailed traffic models and new intelligent transportation systems, which all rely on the understanding of route choice behavior of ideally all traffic participants. Enhanced transportation systems could potentially minimize noise and pollution and save energy and thus are a relevant part of a smart city (Chen, Ardila-Gomez, and Frame 2017). Second, floating car data (FCD), which is data revealing actual paths of drivers and is often used in route choice studies, has recently become more available for research purposes. Older studies often had to rely on stated preference data, which leads to more uncertainty as planned or stated routes often deviate significantly from those that are actually chosen in the end (Papinski and Scott 2013). Third, better capabilities to deal with big data for example through the power of the cloud and through sophisticated analytics and advances in database management systems, have led to new ways in which route choice behavior is analyzed. For these reasons, it has now become possible to test hypotheses about route choice behavior using a vast number of observed paths.

In route choice studies the behavior of traffic participants ranging from pedestrians (Hoogendoorn and Bovy 2004) over bicyclists (Menghini et al. 2010) and motorists (Ciscal-Terry et al. 2016) all the way to truck drivers (Knorring, He, and Kornhauser 2005) has been analyzed. This work focuses on the analysis of taxi drivers. Understanding the nature of taxi drivers' route choice behavior is not only essential for traffic modeling (Sun, Karwan, and Kwon 2016) and the development of intelligent transportation systems (Sun et al. 2014) as already mentioned above, but studying how taxi drivers make route decisions will also provide important insights to improve existing car navigation systems, which so far mostly provide shortest or fastest routes. By analyzing millions of observed paths of taxi drivers in Vienna using GPS-based floating car data, this study's aim is a detailed investigation of the route choice behavior of taxi drivers when having passengers onboard. More specifically, the following questions are the focus of the present work: Do taxi drivers with passengers onboard take the fastest or shortest routes? How do the chosen routes differ from fastest and shortest routes? What are the route characteristics preferred by taxi drivers with passengers onboard? How do certain external factors affect the route choice behavior? How do they choose their routes in certain scenarios? What is the influence of network centrality and geography? Ultimately, the study also checks to some extent, whether two existing route choice models – the shortest path and the anchor-based model – are adequate for describing the behavior of taxi drivers.

The analysis consists of three main parts: First, observed paths are compared to shortest distance and fastest free-flow travel time paths with the use of eight variables describing a route's characteristics. Second, observed paths are split into groups to assess what influence three external factors, namely time of the day, day of the week and weather conditions, have on route choice behavior. Last, taxi drivers' behavior is investigated in more detail with the use of case studies. In these, the effects of network centrality and geography are analyzed. Furthermore, the influence of changing external factors on the location of chosen paths is examined and it is investigated if direction of travel has an influence on route choice behavior.

This work is structured as follows: First, the next chapter discusses and summarizes the related work relevant for this study. It gives an overview over the big academic field of route choice and shows how floating car data has been used in the literature so far. Additionally, the most relevant papers about route choice behavior of taxi drivers are highlighted. This chapter also includes a brief review of the literature on network centrality measures. It is important to look at these since they are used in several parts of this study. The second chapter ends with the identification of some research gaps and open questions. In chapter 3, this thesis is motivated and the research objectives and questions as well as some initial hypotheses are outlined. Chapter 4 gives an overview of the methodology of this work. This includes a description of the data and tools that have been used during the analysis stage. The whole workflow is delineated in detail and some limitations of the chosen approach are discussed. The results of the analysis are displayed in chapter 5 and subsequently discussed in chapter 6. This paper concludes with chapter 7, which summarizes the most important findings and outlines future research possibilities and lists some open questions.

2. Related Work

This work is embedded in the field of route choice research. To efficiently analyze and eventually understand route choice behavior of taxi drivers, it is necessary to get an overview of the academic field first. Therefore, this chapter begins with an overview over the existing literature. It starts with a more general introduction and after that follow more and more detailed and specific discussions of papers, theories and findings as they get more relevant for the present study. First, following Prato's summary of the field (Prato 2009), a brief introduction to route choice modeling is given. Starting from that, different branches of the field are identified, and their most important works are presented. We will see that there exist two broadly distinct approaches, the behavioral approaches and approaches stemming from transportation studies (Manley, Addison, and Cheng 2015). Next, two approaches of route choice research that are highly relevant for the present study, are discussed: The shortest-path approach and the anchor-based approach. The former is frequently used in studies in the field while the latter is relatively new and presently less dominant. Both approaches are investigated during the analysis part of this work. Following Manley, Addison, and Cheng (2015), this study seeks to determine if these models are valid for the case of taxi drivers in Vienna and to see if they can explain the route choice behavior effectively. After this overview over the different approaches and underlying models, the discussion then turns to some methodological differences and other relevant aspects of this thesis are listed, such as the influence of spatial knowledge and experience on route choice behavior. This first part of the chapter ends with an overview of research that has been conducted with taxi floating car data.

An introduction to network centrality indices forms the second part of this chapter. Centrality indices are an important part of this work as they are used in a multitude of ways during the analysis of route choice behavior of taxi drivers. The centrality indices relevant for the present work are degree, closeness, betweenness and information centrality. Thus, the definition of these four indices is included in this chapter. Additionally, the meaning of these indices and their application for route choice research and some related areas will be discussed.

Taking these summaries of the existing bodies of literature as a starting point, the identification of open questions and research gaps makes up the last section of this chapter. This then leads to chapter 3, in which the research questions for the present work are outlined.

2.1 Route Choice

Route choice studies have been around for quite some time now. One of the earlier and most cited studies is Wardrop's paper "Some Theoretical Aspects of Road Traffic Research" from 1952. In his paper, Wardrop says that "the journey times in all routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route". This is known as Wardrop's first principle. To this day, this assumption has been questioned by many researchers (Zhu and Levinson 2015; Manley, Addison, and Cheng 2015) and new models explaining route choice behavior have been developed. Generally, there is now an abundance of papers on route choice behavior and it is necessary to classify and distinguish these studies to get an overview of the field and to position the present thesis in it.

There exist many ways in which this could be done. There are scholars from many different fields ranging from transportation research over psychology to behavioral geography doing research on route choice (Manley, Addison, and Cheng 2015; Venigalla, Zhou, and Zhu 2017), which has led to many different approaches and methodologies for the analysis of the behavior of many different groups of traffic participants. In this chapter, these various branches of route choice research are identified and discussed. Hereby, more attention is given to papers which are recent and closely connected to the present work. After that some methodological differences are outlined.

2.1.1 Route choice modeling

Route choice research is often also called route choice modeling (but does not necessarily always refer to the same thing) as the aim of most studies in route choice is to come up with a model that captures the behavior of drivers. These models can then be used to predict future traffic volumes or assess the influence of changes in a road network on traffic flow. All choice models aim to predict which route out of all possibilities travelers choose in the end. Additionally, they try to explain how and why a route is chosen (Prato 2009). Route choice modeling may be divided into two broadly distinct fields. These fields are transportation studies on one hand and spatial cognition and behavioral geography research on the other hand. Transportation studies focuses more on how traffic distributes around the road network and on the relationship between drivers and roadway engineering (Manley, Addison, and Cheng 2015). Here, the underlying approach of most studies is to first generate a so-called choice set, which consists of all feasible paths from an origin to a destination. The second stage consists of the creation of a model, which predicts which routes out of this set will be chosen by drivers in reality (Prato 2009). These models can then be calibrated, validated and verified by evaluating stated preference data or floating car data. Stated preference data is collected by asking drivers to describe what route they would choose in certain scenarios. This is usually done in an experiment-like setting where people point out their preferences on a map. In contrast, floating car data shows the choices that drivers actually made, as this data is collected during real trips by a GPS-device. Irrespective of these differences in terms of model and data, the overall approach is often very similar and will now be discussed in more detail.

Choice set generation

First, a choice set must be generated. The aim of this is to identify the so-called consideration set. The consideration set contains all routes, which a specific person considers as viable paths between a specific origin and destination-pair. In the end, route choice models try to explain, which route out of this consideration set will most likely be chosen by the traveler. There are two steps to get to the consideration set (see Figure 1). First, a master set has to be created. The master set is a subset of the universal realm, which contains all possible paths between an origin and destination. There is usually an enormous number of potential paths even in a mildly complex street network, thus it is impossible or at least impractical to generate all paths of the universal realm. Additionally, most paths will not be considered by anyone as they contain large detours or loops. Therefore, it is necessary to use route generation techniques to approximate the set of routes known to the traveler (Kaplan and Prato 2012). Such a technique could for example be a k-shortest path-algorithm, which creates *k* paths that minimize a certain attribute of a route, such as travel distance or the number of turns. The size of the master set naturally varies from traveler to traveler and depends on the spatial knowledge of the driver (Prato 2009). The next step is to get from the master set to the consideration set. This is done by incorporating personal preferences, characteristics and network familiarity of the traveler to generate all routes, which the driver considers when planning a trip (Kaplan and Prato 2012). These steps make up the first half of Figure 1, the choice set formation.

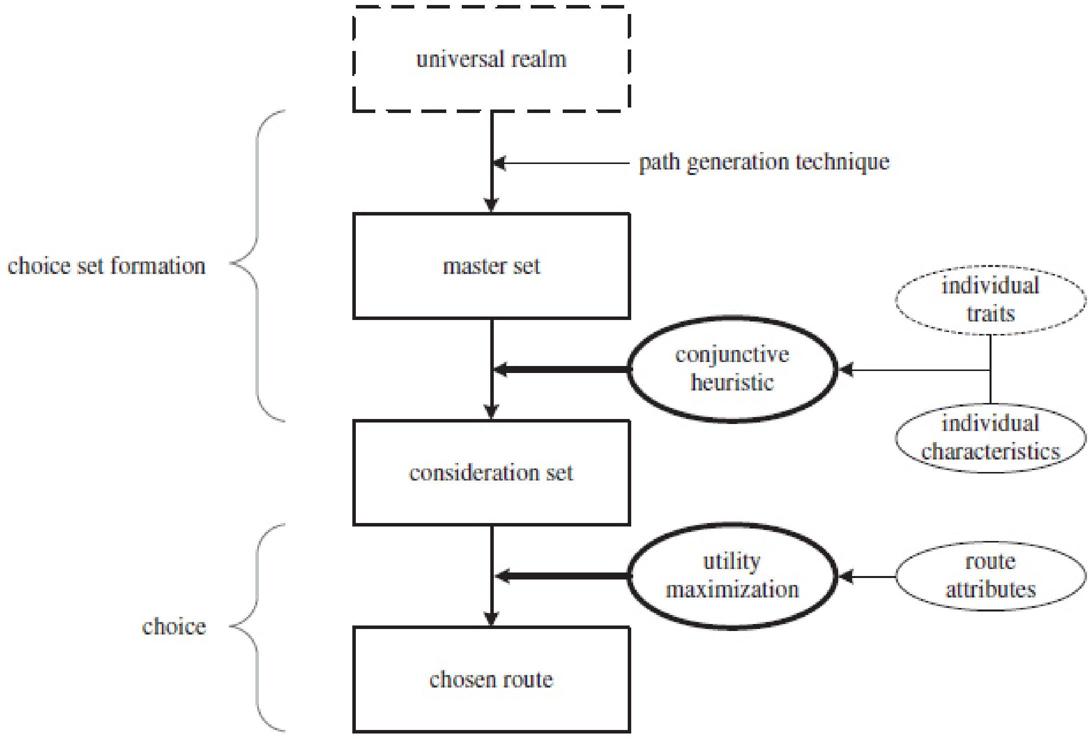

Figure 1: Process of choice set formation and choice from the universal realm to the chosen route (Kaplan and Prato 2012).

Choice process

The second part of route choice modeling is the analysis of the choice process itself (see second part "choice" in Figure 1). The key question is: Which routes from the consideration set are travelers going to choose and, more importantly, why do they choose these particular routes? To answer these questions, researchers came up with an abundance of choice models. These can be categorized in many ways. Prato (2009) divides choice models into models using a Logit structure, a GEV (Generalized Extreme Values) structure or a non-GEV structure. It is outside the scope of this thesis to define and discuss these models. However, it is important to note that recent models assume that drivers maximize their utility. Earlier on, for example in Wardrop (1952) models were often based on the assumption that individuals optimize their travel time between origin and destination (Manley, Addison, and Cheng 2015). Studies have shown that assumption is not valid for most individuals as it assumes unrealistically that travelers have perfect knowledge about path costs and choose the route that minimizes their travel costs (Prato 2009; Manley, Addison, and Cheng 2015; Garling, Laitila, and Westlin 1998). Other studies have found out that there also exist strong intrapersonal variability in route choice behavior (Pas and Sundar 1995). Thus the characteristics, preferences and experiences of drivers have a strong effect on the outcome of the choice process. Also, studies have shown that latent variables, such as gender, age and education also affect the behavior of drivers (Prato, Bekhor, and Pronello 2012; Papinski and Scott 2013; Ramaekers et al. 2013).

2.1.2 Minimum Cost Path Approaches

Now we look a bit more closely at two other route choice modeling approaches, both of which are part of the present thesis. The minimum cost path approach is based on the assumption that travelers try to minimize one or more route characteristics (Zhu and Levinson 2015). Examples for such characteristics are number of turns, travel distance, travel time, number of intersections along a route or route directness. From these, travel distance and travel time have received the most attention in the literature (Papinski and Scott 2011). Table 1 gives an overview over some variables that have been analyzed in several studies.

Table 1: Analyzed route characteristics in selected studies.

Study	Route Characteristics
Papinski and Scott (2011)	A, C, D, E, F, G, H, I
Manley, Orr and Cheng (2015)	B, C, I
Ciscal-Terry et al. (2016)	C, B, G, I
Zhu and Levinson (2015)	A, B, C, J, K
Venigalla, Zhou and Zhu (2017)	B, C, D, F, L
Sun et al. (2014)	H
Kaplan and Prato (2012)	A, C, D, F
Ramaekers et al. (2013)	C, F
Papinski and Scott (2013)	G

A	Free-flow travel time
B	Estimated travel time
C	Travel distance
D	Number of turns
E	Number of links
F	Road type
G	Route Directness Index
H	Percentage of shared length
I	Speed
J	Maximize freeways
K	Minimize freeways
L	Number of signals

Regarding travel distance, Jayasinghe et al. (2016) differentiate between metric, topological and geometric distance. Metric distance is simply the length of a route in meters or kilometers. Topological distance is defined as the number of turns of a route. The more turns, the longer the topological distance. Lastly, geometric distance is defined as sum of angles of all turns of a route. Generally, they showed that none of these three metrics can fully explain the behavior of drivers. However, ordinary drivers in cars seem to be rather minimizing geometric distance than topological or metric distance. In contrast, taxi drivers seem to give more weight to the metric distance relative to the other two metrics. Generally, shortest distance paths show a low overlap with observed paths: The values for percentages of similarity range from 14 percent (Sun et al. 2014) over 35 percent (Venigalla, Zhou, and Zhu 2017) to 39 percent (Manley, Addison, and Cheng 2015), which is rather low. However, it is important to note that the percentage of overlap heavily depends on the Euclidean distance between origin and destination. In cases where origin and destination are close to each other, the overlap of observed and minimum distance paths is relatively high. This percentage then drops rather quickly as trips get longer, while for very long trips, the percentage increases again (Zhu and Levinson 2015; Venigalla, Zhou, and Zhu 2017). Additionally, there is a difference between total overlap and partial overlap. For example only 14 percent of all routes analyzed in Sun et al. (2014) have a total overlap with shortest distance paths. However, 53 percent of all routes have an overlap of 50 to 80 percent with their respective minimum distance paths.

As for shortest travel time paths, there exist more ways in which these can be created than for shortest distance paths, as travel time cannot be derived directly from a street networks' geometry without any additional information. Rather, the calculation of travel times can either be based on free-flow travel time derived from speed limits or on time-dependent average trip speeds extracted directly from the FCD data at hand. An example of the latter option can be found in Ciscal-Terry et al. (2016). The authors calculated average road speeds by summing up recorded speed values of trips grouped by road segments for every hour of the day. Segments that did not have any speed information recorded on them, have been assigned an average value of adjacent segments. Irrespective of the way travel times are computed, minimum travel time paths also do not have very a high overlap with observed paths, thus performing badly when trying to predict chosen routes. Values for the percentage of trips overlapping with their minimal travel time paths range from to 16 percent (free-flow) and 23 percent (observed; Zhu and Levinson 2015) over 22 percent (free-flow; Venigalla, Zhou, and Zhu 2017) to 38 percent (observed travel times; Manley, Addison, and Cheng 2015).

Other, more advanced models exist, which do not include a single variable, but combine multiple ones, to predict routes, that overlap better with observed paths. However, none of these captures the behavior of all drivers perfectly (Manley, Addison, and Cheng 2015) and thus, chosen routes from a specific origin to a destination often deviate significantly from hypothetical optimal ones.

Most studies then compare the observed routes with their corresponding shortest or fastest paths not only in terms of overlap but also by trying to describe differences in route characteristics between computed and observed paths. This comparison usually focuses on one or more of the following aspects: travel distance, travel time, number of turns, road class (highway, local road, etc.), route directness (angular deviation between the straight line distance between origin and destination and the actual route's distance), number of road segments, number of signals along the route, route complexity and number of intersections (see Table 1). Table 2 shows what models have been assessed in selected papers.

Table 2: Analyzed route choice models in selected studies.

Study	Route Choice Model
Papinski and Scott (2011)	A, B
Manley, Orr and Cheng (2015)	A, C, F, G, H
Ciscal-Terry et al. (2016)	A, C, I, J
Zhu and Levinson (2015)	A, C
Venigalla, Zhou and Zhu (2017)	A, C
Manley, Addison and Cheng (2015)	A, B, C, K, L, M
Gao et al. (2013)	D
Sun et al. (2014)	A, C
Zhao, Zhao and Cui (2017)	N, E
Papinski and Scott (2013)	A, B

A	Minimal distance
B	Minmal free-flow travel time
C	Minimal estimated travel time
D	Maximal betweenness centrality
E	Maximal degree centrality
F	Least total deviation from gateway
G	Maximal mean speed
H	Least distance from target
I	Shortest path with turn penalties
J	Shortest path with traffic lights penalties
K	Minimal angular deviation
L	Fewest turns
M	Fewest road segments
N	Maximal PageRank centrality

Influence of factors

Some studies (e.g. Jayasinghe et al. 2016; Papinski and Scott 2011) then go a step further and look at how certain external factors, which usually cannot be derived directly from the geometry of the trajectories, influence the characteristics of observed routes in comparison to computed ones. This is done by splitting the routes into separate groups to find out what influence factors - such as trip length, time of the day, day of the week, and weather conditions - have on the chosen routes' characteristics. Four of these shall briefly be discussed below:

1. Length

Route length has a clear influence on route characteristics as has already been described above for the percentage of overlap between shortest distance and observed paths. Venigalla, Zhou, and Zhu (2017) report differing results for other route characteristics depending on the length as well: The longer a trip, the fewer signals and turns there are per mile along a route. Furthermore, the share of primary roads gets higher with increasing trip length, while the share of tertiary roads decreases.

2. Weather

The influence of weather conditions is less clear but has also received some attention in the literature (Liu, Susilo, and Karlström 2017). However, most findings concern the influence on demand and supply for transportation systems and not on route choice behavior itself. Nevertheless, there are some studies which found out how weather affects the behavior of drivers: Adverse weather (rain and snow) makes drivers go slower, thus such conditions generally have a negative impact on traffic speed (Edwards 1999; Oh, Shim, and Cho 2002; Akin, Sisiopiku, and Skabardonis 2011). This is despite reports saying that there is a negative correlation between rain or snow and the amount of traffic (Keay and Simmonds 2005; Datla and Sharma 2010). To conclude, even though there is less traffic under adverse weather conditions, people seem to drive slower and more carefully.

3. Time

Time is likely to have a considerable influence on route choice behavior of drivers and thus also on chosen routes' characteristics. There are several papers, which confirm this assumption. Yang et al. (2017) show that time of the day of a trip has a strong influence on route choice and therefore, they suggest taking these factors into account when trying to extract knowledge from taxi trajectory data for the enhancement of navigation systems. Manley, Addison, and Cheng (2015) report variations between trips depending on the time of travel, however, they were smaller than the variations between differing trip lengths.

4. Geography

There is one factor, which has received surprisingly little attention from scholars: Geography. The route choice behavior as well as route characteristics differ considerably depending on the location and geography of the area that the trip go through (Ciscal-Terry et al. 2016; Yang et al. 2017). In a case study in the Reggio Emilia territory in Italy, Ciscal-Terry et al. (2016) found out that the heterogeneity of route choices is higher in plain regions than in mountainous areas. They conclude that thus location – because of differing road infrastructure configurations and properties – has a significant effect on route choice behavior.

For the present work, the above factors are classified as either internal or external. Length is the only internal factor as this property can directly be derived from the trajectory of a trip. The other three factors are external ones as these are not geometric properties of a trip. In this study, only the influence of external factors is assessed. Table 3 lists selected studies that have analyzed one or more of the four factors described above.

Table 3: Analyzed factors that influence route choice behavior in selected studies.

Study	Factors
Venigalla, Zhou and Zhu (2017)	C
Yang et al. (2017)	A
Manley, Addison and Cheng (2015)	A, C, D
Ciscal-Terry (2016)	D
Edwards (1999)	B
Oh, Shim and Cho (2002)	B
Akin, Sisiopiku and Skabardonis (2011)	B

A	Time of day
B	Weather
C	Route length
D	Geography

2.1.3 Anchor-based approach

Aside from the studies described at the beginning of this chapter, which all follow a similar approach – they often only look at the road network and personal characteristics and follow the utility maximization paradigm – and the shortest path approach, which has received quite a lot of critique (Zhu and Levinson 2015; Manley, Addison, and Cheng 2015; Ciscal-Terry et al. 2016), there exist some interesting papers which look at the route choice process in more detail. The approach introduced here is called the anchor-based approach and is based on a more individual-centric perspective. It stems from the fields of behavioral geography and spatial cognition and has been proposed by Manley, Addison, and Cheng (2015). They reject the idea that people construct their routes at once and, more importantly, they show that individuals do not seem to follow any kind of shortest paths (nineteen shortest routing mechanisms were considered in the study, all of which performed quite poorly). Thus, they came up with this new framework which assumes that drivers plan their routes around anchors or landmarks.

Heuristic model of route choice behavior

One key feature of the anchor-based approach is that space is modeled with a hierarchical representation of space in mind (Manley, Orr, and Cheng 2015). In contrast to other frameworks, it does not rely on simplistic assumptions around the nature of human cognitive ability, memory and preference. Rather, by incorporating a hierarchical model of space and acknowledging the uncertain nature of route choice decision making, the authors describe heuristic rules under which travelers make route choice decisions. More importantly, route choices are not made instantly, but rather are a result of a multi-step process where locations and urban features play a key role (Manley, Addison, and Cheng 2015). One of the authors' main points of critique of existing route choice model frameworks is the assumption that people have unlimited spatial knowledge and make their decisions under conditions without any forms of uncertainty. Additionally, the authors claim that people usually do not try to find an optimal solution – e.g. the "best" path - but rather seek to come up with a "good enough solution" in a considerably brief time. More interestingly, they provide another perspective on urban space. While many traditional route choice studies are merely based on the road network, the authors – drawing from research from neuroscience, cognitive science and behavioral geography – assume a much more complex relationship between individuals and the urban space. People seem to use landmarks as reference points and their mental representation of urban space is hierarchical. Therefore, the novel approach uses a hierarchical representation of space which recognizes that navigation decisions in urban areas may be based on various spatial features, rather than purely road connectivity and route conditions. The urban space is modelled using regions, nodes and roads, where these three feature types form the hierarchy (see Figure 2). Every feature type is connected to a step in the decision-making process. First, regions are used as the basis for an initial coarse route plan, which is then refined during the subsequent steps. Regions are parts of a city, which are distinct from each other in some way. Therefore, regions have a lot in common with one of the five elements of a city from Lynch (1960), namely districts. Furthermore, regions relate to each other by gateways (see Figure 2). Regions contain a collection of nodes, which are used for the next step: They connect the regions from the previous step in a more refined way. More specifically, routes are constructed by connecting different nodes, which represent distinct urban features, such as landmarks or big intersections. Third, roads, which are at the lowest level of the hierarchy, are used to connect one node to another until the final destination has been reached.

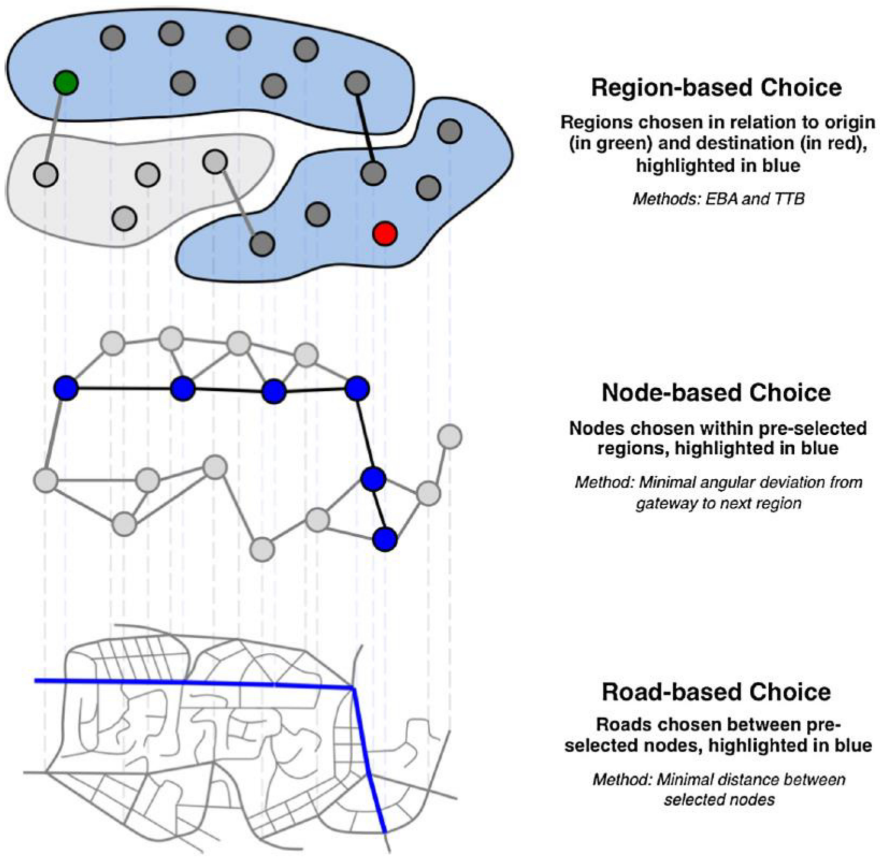

Figure 2: Hierarchy of urban space and their corresponding route choice steps (Manley, Orr, and Cheng 2015).

This framework has been tested by Manley, Addison, and Cheng (2015) by analyzing a large-scale dataset with floating car data from mini-cab drivers in London. They found quite a lot of evidence for the validity of the anchor-based approach. They showed that in a lot of case studies, a majority of routes from one post-code area to another seem to follow along anchors, which have been identified as mostly being big intersections. Furthermore, the authors were able to identify a lot of directional effects in the data, meaning that in many scenarios, drivers choose different routes from A to B than from B to A (A and B being some origin/destination areas in London based on postal codes). From this, they concluded that these directional effects are due to anchor-based route choice of drivers and that they plan their routes around those while driving.

2.1.4 Navigation, wayfinding and the role of experience

Next, another body of literature is briefly introduced: navigation and wayfinding. Authors in these two related fields bring in another perspective on how people plan and choose their routes. One of the main aims of many studies in these areas is to find out, how people acquire spatial knowledge and how this affects their capabilities to find their way around a city for example. Stern and Leiser (1988) and others identify three stages of spatial knowledge: landmark, route and survey knowledge. In the landmark stage, people can identify individual important locations, but are unable to link them to each other in space. In the second stage, people can link landmarks to each other and thus find their way around by linking routes connecting landmarks. However, they do not have a full knowledge of the configuration of space yet, their mental map is distorted, and they do not know many alternatives from an origin to a destination. When progressing to the final stage, people gain survey knowledge, which means that they have a proper knowledge about the configuration of the space in an area and can create complete routes linking origin and destination beforehand and do not need to plan their trips around landmarks involving temporally ordered route choices anymore (Montello 1998; Wiener, Büchner, and Hölscher 2009; Montello 2005). In this context it is interesting that directional effects can also be linked to these different stages and not necessarily to anchors. Moar and Carleton (1982) found out that drivers with less experience and spatial knowledge show a strong directional bias, meaning that the chosen routes in one direction linking two specific locations are asymmetric compared to chosen routes in the other direction. Such directional effects were absent for drivers with survey knowledge. More importantly, several studies indicate that route choice differs significantly between professional and non-professional drivers (Stern and Leiser 1988; Stern and Bovy 1990; Manley 2016; Pailhous 1970). More specifically, the number of known as well as the number of selected routes vary according to these two variables (see Figure 3). Additionally, professional drivers clearly select more optimal routes than non-professionals (see Figure 4). The authors also suggest that non-professional drivers will never know as many alternative routes from A to B as professionals. Non-professionals usually know one or two alternatives, which is sufficient for their needs. Learning a lot of additional alternatives and learning to know which of them are best under varying conditions takes a lot of effort. Only professional drivers are motivated to do so and optimize their route selection (Stern and Leiser 1988).

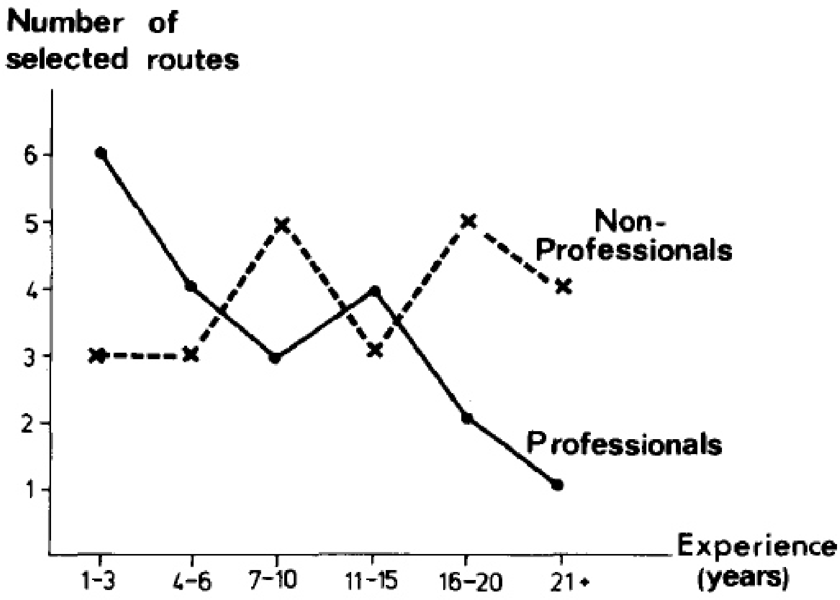

Figure 3: Changes in the number of selected routes with increasing driving experience (Stern and Leiser 1988).

Evidence for the superior spatial knowledge of taxi drivers can also be found in the results of study from Maguire et al. (2000). According to their findings, there are structural changes in the hippocampus of experienced taxi drivers compared to ordinary drivers. These areas of the hippocampus that are bigger in the case of taxi drivers seem to be involved when previously learned spatial information is used and during the encoding of new environmental layouts. Taxi drivers' representation of the city they operate in (in this case London, UK), is reported to be more extensive than that from ordinary drivers. The mental map of taxi drivers of the city gets more accurate over time, permitting them to relate routes and places to each other.

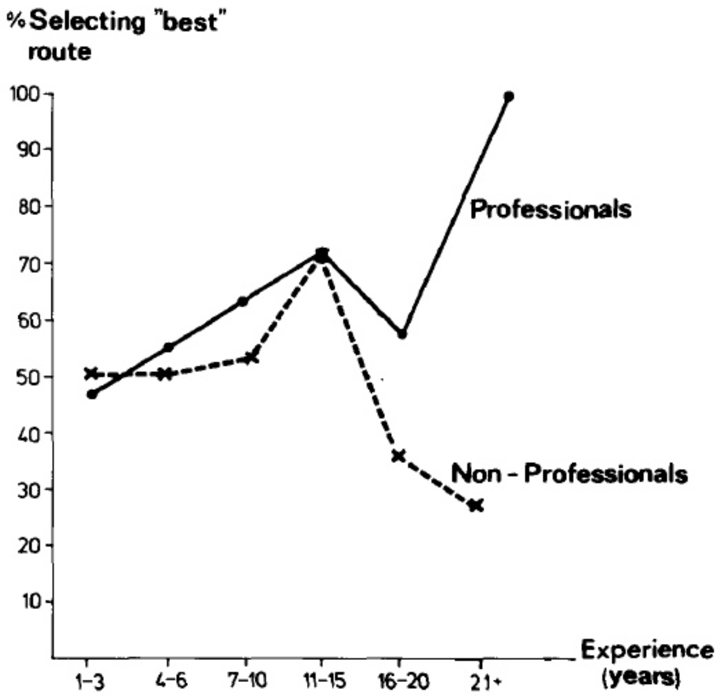

Figure 4: Differences in selecting optimal routes between professionals and non-professionals (Stern and Leiser 1988).

2.1.5 Methodological differences

Route choice research also differs a lot depending on what data and methodology is used. Here, three often used approaches are briefly described highlighting their differences and individual strengths and weaknesses.

Stated preference method

A lot of older studies had to rely on stated preference data, where for example traffic participants were asked what route they would choose in a given scenario, or the study participants had to describe their home-to-work commute. While stated preference data has been and still is being used successfully, this methodology has some drawbacks. For example, people may recall a route incorrectly or people state they would choose a route that they think is favored by the researcher, while in reality they would choose another route (Papinski, Scott, and Doherty 2009).

Floating car data research

More recently, a lot of studies made use of floating car data, which enables to capture the behavior of drivers directly (Papinski and Scott 2011). Usually the FCD is map-matched first to generate trajectories on a street network. Subsequently, these trajectories can be analyzed to investigate the route choice behavior itself. Thus, a lot of data cleaning and preprocessing is usually involved when using FCD, but in the end, this kind of data offers many ways to analyze route choice behavior in great detail.

Case studies

There are a few papers that investigate case studies to analyze the route-choice behavior of motorists in more detail and that not just rely on the analysis of route variables without looking at the location of observed trips in space: While Manley, Addison, and Cheng (2015) use multiple case studies to better understand the route choice behavior of minicab drivers in London, Thomas and Tutert (2015) try to find out if people prefer a central or an orbital route, when they have to go from one end of the city of Enschede, Netherlands to the opposite end. One weakness of this case study approach is that it is difficult to verify any hypothesis when only using case studies. As long as not every case has been investigated, the possibility that other cases would contradict the findings of earlier case studies cannot be dismissed. The advantage is that one can investigate route choice behavior in much more detail. For example, the identification of landmarks and linking them to route choice behavior as has been done in Manley, Addison, and Cheng (2015), is only possible with case studies. Additionally, the influence of geography and location can be captured best using this approach. Looking at trips in space is important as considering route characteristics only cannot capture all aspects of the route choice behavior of drivers.

2.1.6 Route choice research of taxi drivers

One of the earliest studies on the behavior of taxi drivers comes from Pailhous (1970), who identified that the mental representation of the street network for Parisian taxi drivers consists of a primary and a secondary network. The primary network consists of major routes, which connect the districts of the city with each other. The secondary network consists of smaller streets within the districts. Thus, a theory is developed that taxi drivers first plan their routes using streets of the primary network to connect origin and destination areas. Within these areas, they recall the secondary network to go to the nearest primary street quickly starting from the point of departure and lastly, reaching the destination as efficiently as possible from the nearest major street. Thus, already this earlier study assumes that taxi drivers choose their routes in a different manner than ordinary drivers and it was noted that they have more spatial knowledge.

Nowadays, browsing through recent papers on route choice, it is evident that not a small number of works make use of floating car data from taxi drivers, which is probably due to the good availability of such data. The present thesis also works with taxi FCD (see chapter 4.1.1 for more details). Because of this and because taxi drivers seem to take more efficient routes than ordinary drivers, there are a couple of studies that try to exploit the knowledge of taxi drivers to enhance navigation systems or enhance traffic models (Yuan et al. 2013; Ji-hua, Ze, and Jun 2013).

Additionally, a lot of research with taxi FCD has been done in areas other than route choice research. For example, Kang and Qin (2016) have investigated taxi drivers' operation patterns by analyzing their digital traces derived from FCD. They looked at how demand and supply for taxis influences the mobility patterns of cab drivers and identified several taxi demand and supply regions. The conclusion is that the operation of taxi drivers is a reaction to city residents' travel activities. Tang et al. (2015) looked at human mobility patterns, again using taxi FCD. They identified popular pick-up and drop-off locations, studied the searching behavior of taxi drivers and constructed a model to estimate the traffic distribution with a large-scale analysis of the trajectories derived from FCD. Similarly, Zhang et al. (2017) looked at the distribution of pick-up and drop-off points and assessed the influence of seasons and day of the week on these distributions. They identified time and season-dependent hot-spots for taxi trips' starting and ending points. Lastly, Graser et al. (2016) combined street network centrality indices and average travel speeds on street network links derived from taxi FCD to improve vehicle speed estimates for the road network in Vienna, Austria.

2.2 Centrality

The notion of centrality that is relevant in the context of the present work, originates from the field of structural sociology. In this field of research, scholars analyze the relationships between individuals in specific networks of people. These networks can be anything from co-workers in an office to all inhabitants of a country (Freeman 1978). To analyze these networks, a graph is created by turning the relationships between people into edges and the individuals into nodes (Porta, Crucitti, and Latora 2006). These graphs can then be subsequently investigated using graph-based analysis methods, such as the centrality indices mentioned earlier. Results from such analyses could for example be the revelation of hidden power structures in a national government, or the description of the way information flows from employees in the lower ranks in a company to the managers, to stay with the initial examples.

A lot of the pioneer work in this field has been conducted by Linton Freeman. In two papers, he defines a set of centrality indices, namely degree centrality, closeness centrality and betweenness centrality (Freeman 1977, 1978). With these indices, it becomes possible to analyze some properties of the social networks such as the relative importance or the reachability of a person. Some additional indices have been introduced since then, such as straightness, PageRank and information centrality. The latter together with the original three indices from Freeman are used in this work and shall therefore be defined now:

Degree centrality

Let $G\ (V,E)$ denote a directed graph, where V is the set of vertices and E the set of edges in the graph. The normalized degree centrality D^{node} of a node i is defined as

$$D^{node}(i) = \frac{\sum_{j \in N} a_{ij}}{N-1}$$

where a_{ij} is the adjacency matrix for node i with all other nodes j in G and N is the number of vertices. Every element in a where (i, j) belongs to E equals to 1. For normalization purposes, the sum of the number of edges between i and all other nodes is divided by the total number of vertices in G subtracted by 1. Therefore, D^{node} takes on values between 0 and 1 and is equal to 1 when a node is connected to all the other nodes of the graph (Porta, Crucitti, and Latora 2006).

Closeness centrality

Normalized closeness centrality is defined as

$$C^{node}(i) = \frac{N-1}{\sum_{\substack{j \in N \\ j \neq i}} d_{ij}}$$

where d_{ij} is the shortest path length between i and j. The shortest path is can be based on any cost. This cost can be for example length (distance between i and j), travel time, or any centrality index. It can also be equal to 1 for every edge in case of a non-valued graph without any attributes. Hence, closeness centrality measures how close a node is to all other nodes in the network (Graser et al. 2016; Porta, Crucitti, and Latora 2006).

Betweenness centrality

Normalized betweenness centrality is defined as

$$B^{node}(i) = \frac{1}{(N-1)(N-2)} * \sum_{\substack{j,k \in N \\ j \neq k; j,k \neq i}} \frac{n_{jk}(i)}{n_{jk}}$$

where n_{jk} is the number of all shortest paths between j and k and $n_{jk}(i)$ is the number of all shortest paths between j and k that contain vertex i. The first term is included for normalization; $B^{node}(i)$ takes on values between 0 and 1 and reaches its maximum when node i falls on all shortest paths. Thus, betweenness measures how central a node is in a given network and how much it lies in between other nodes in terms of shortest paths (Porta, Crucitti, and Latora 2006; Graser et al. 2016).

Information centrality

Information centrality is defined as

$$I^{node}(i) = \frac{\Delta E_{glob}}{E_{glob}} = \frac{\Delta E_{glob}(G) - \Delta E_{glob}(G')}{E_{glob}(G)}$$

where E_{glob} is defined as

$$E^{glob}(G) = \left(\sum_{i,j \in N; i \neq k} \frac{d_{ij^{Eucl}}}{d_{ij}} \right) \Big/ N(N-1)$$

where $d_{ij^{Eucl}}$ is the Euclidean distance between nodes i and j along a straight line and d_{ij} is again the is the shortest path length between i and j. E_{glob}defines the efficiency of a graph. When a node i is removed from the graph, and i is part of a lot of shortest paths in the graph, the efficiency of the latter decreases. More specifically, Porta, Crucitti, and Latora (2006) state that "the information centrality of a node i is defined as the relative drop in the network efficiency caused by the removal of the edges from G incident in i.". Information centrality thus indicates how critical a node is in a specific network. Therefore, when the global efficiency decreases a lot when all edges incident in i are removed, then the information centrality value of i is high.

2.2.1 Centrality of street networks

In the fields of geography, GIScience and route choice, centrality indices have been applied to street networks in various manners. Liu et al. (2018) classified the bicycle road network of Salt Lake City, USA into different road categories (connector road, peripheral road etc.) using two variants of betweenness indices and a single degree centrality index. The results showed that these indices are a good indicator of the attractiveness of a distinct road for bicyclists.

As already mentioned before, Graser et al. (2016) used centrality indices to enhance speed estimates on road links. They used betweenness and closeness centrality to do so and found that closeness centrality helps to distinguish between central urban and peripheral rural streets and that betweenness centrality helps to identify important links in a street network.

Tomko, Winter, and Claramunt (2008) created a street hierarchy derived from the betweenness centrality values of the individual roads. They argue that betweenness centrality reflects the hierarchical importance of streets in the city network. Additionally, they claim that closeness centrality fails to do so.

One challenge for the appliance of centrality indices that has gained quite some attention among scholars concerns the way the street network is represented and thus, subsequently, how the centralities are computed. There exist two main approaches for that: the primal approach and the dual approach. In the dual approach a graph is created in which streets are turned into nodes and intersections are turned into edges. In contrast, In the primal approach it is exactly the other way around: intersections are turned into nodes, streets are turned into edges (Porta, Crucitti, and Latora 2006). The dual approach stems from the field of space syntax which is a term prevalently used by scholars of the field of architecture. For the analysis of street networks, the primal approach is reported to be more suitable, because it "allows a metric computation of distance without abandoning the topology of the system, the dual-generalized approach leads to only a topological computation of distance, which makes indices and processes fundamentally more abstract, in the sense that they appear to miss a relevant part of the causal factors of collective behaviors in space." (Porta, Crucitti, and Latora 2006). Therefore, in this study a primal representation is chosen for the road network of Vienna.

Another issue is that some centrality indices suffer heavily from border or edge effects, where streets in the center of the study area generally receive higher values than streets at the borders of the area (Tomko, Winter, and Claramunt 2008; Porta, Crucitti, and Latora 2006). Especially closeness and information centrality are prone to these effects, which has lead scholars to also compute local centrality indices, to mitigate border effects (Graser et al. 2016). However still, there are concerns about whether centrality indices can adequately capture the importance of streets. For example, the measure of closeness centrality fails to reveal the relative importance of the streets to the overall structure. Betweenness centrality seems to perform better for this task (Tomko, Winter, and Claramunt 2008).

2.2.2 Centrality as predictors of traffic flows

Several studies have used centrality indices to analyze route choice behavior and to predict traffic flows. Zhao, Zhao, and Cui (2017) used enhanced centrality indices (degree, betweenness and PageRank) considering both the topological characteristics and the geometric properties of the road network to analyze and predict traffic flows in Wuhan, China. They concluded that these modified measures have higher correlation with observed traffic flows than the ordinary ones. in a similar approach, Gao et al. (2013) predicted traffic flows using betweenness centrality. They found that betweenness centrality together with some constraints (distance-decay law and spatial heterogeneity of human activities) can explain traffic flows adequately.

2.3 Research gaps

To sum up, there are quite a few studies that are have analyzed the route-choice behavior of motorists. However, there is still room for further research in this area:

First, most studies look at the behavior of ordinary drivers by analyzing commuting or shopping trips. Thus, the investigation of professional drivers, namely taxi drivers could lead to interesting findings. Taxi drivers usually have good knowledge of the road network of the city in which they operate and they are generally very experienced drivers (Pailhous 1970; Yuan et al. 2013). While others might also be experienced, the average driver probably drives the same routes most of the time (e.g. from work to home). Taxi drivers on the other hand make a lot of unique journeys that do not repeat themselves very often, rendering the analysis of their behavior very interesting. The cab drivers in the Manley studies (Manley 2016; Manley, Orr, and Cheng 2015; Manley, Addison, and Cheng 2015) are not licensed cab drivers and therefore probably have less knowledge about the street network than licensed drivers which have to pass a test to obtain a taxi license. For London for example, aspiring cab drivers sometimes have up to four years of training before getting a license (Woollett and Maguire 2010). Therefore, a similar analysis for the case of taxi drivers from Vienna could reveal differences in the route choice behavior of licenses taxi drivers and minicab drivers that presumably have less experience and knowledge. It is still unclear if taxi drivers choose their routes iteratively while driving, or if they have the complete route in mind beforehand.

Second, the lack of case studies and – related to this – the incorporation of geography into the analysis is evident when looking at the papers discussed here so far. Looking at differences in route choice behavior in different regions of a city is promising for the exploration of drivers' choices. With the use of case studies, one can also look at how chosen paths deviate from computed paths in space, which adds another measure for the comparison of groups of paths. More importantly, case studies can also be used to identify areas that drivers tend to avoid as well as identify streets that drivers prefer. Once these have been identified, hypotheses can then be created that explain these observations.

Third, the incorporation of certain factors, especially those that cannot simply be derived from the geometry of trajectories (such as weather, time of the day, day of the week, location of both origin and destination) of the road network is not very common. Additionally, the relative importance of these external factors on route-choice behavior is still unclear is not entirely investigated yet (Zhu and Levinson 2015). This is especially true for the factor of weather conditions. Again, the use of case studies to look at the influence of these factors, could help to identify new patterns, which would otherwise remain hidden when only looking at raw numbers for route characteristics. By applying the spatial clustering method described in Manley, Addison, and Cheng (2015), it is possible to check if there are also clusters for different weather conditions, different times of the day and different days of the week. So far, this method has only been applied to look for directional effects but not for other factors to the best of my knowledge.

Fourth, the influence of centrality and complexity of the road network on the route-choice behavior of drivers has not yet been fully investigated. Most of the time, centrality indices in the literature are only used to compute paths that maximize centrality of a route. This is just another variant of a shortest path model, where the aim is to predict chosen routes. However, there are still some additional ways to link road network centrality and route choice behavior, which have not been explored in detail so far. The street network of a city could be divided into areas with low centrality values and areas with high values. Afterwards, it becomes possible to investigate, if there is a relationship between observed paths and these areas. A potential hypothesis could be that drivers tend to avoid certain areas with a complex street network with high information and closeness centrality values. Furthermore, using centrality indices, geography can be incorporated into the analysis. A city can be divided into several distinct areas based on the average centrality of the streets within these areas. Subsequently, the influence of geography can then be assessed by comparing trips in terms of route characteristics within these different areas. Centrality comes in to play in three different manners in the present work:

1. As a route characteristic. Through the calculation of the centrality of all paths in the database, it is possible to check if there are significant differences in terms of centrality values.
2. As a spatial unit of analysis. It will be investigated if paths in regions with a high degree of centrality show distinctive characteristics that differ to those in regions with a low level of centrality.
3. As the basis for some case studies. It will be analyzed if people prefer regions with low or high degrees of centrality and if drivers avoid certain regions, and if these can be connected to centrality.

Last, it is always good to analyze the behavior of drivers in different cities and in new contexts. Thus, the application of existing techniques to a new city helps to confirm (or reject) findings of other studies.

After the summary of the related work and the identification of some research gaps, the next chapter defines what research questions will be addressed. Afterwards, the data and the methodology of the thesis are described in more detail.

3. Research objectives

The aim of this paper is to analyze chosen routes of taxi drivers in Vienna in detail, assess the influence of time of the day, day of the week and weather conditions on these characteristics, and last, gain some new insights into the route choice behavior of taxi drivers mainly with the use of detailed case studies. This work does not focus on intra- or inter-personal differences as such information is not present in the data at hand. Also, no model is created, which predicts actual routes as good as possible and no route choice models other than the shortest path- and the anchor-based model are incorporated into the analysis.

3.1 Research questions

The following main research questions will be addressed: (1) What are the characteristics of routes that taxi drivers choose, and how do they differ from shortest and fastest routes? (2) What is the influence of certain external factors on these differences and how do the chosen routes differ in space when an external factor varies? (3) Are there any directional effects? (4) Is there any connection between network centrality and route choice behavior?

Ultimately, the aim is to gain some insight into the way taxi drivers choose their routes. Do they plan them beforehand? Do they try to minimize any route characteristic (distance, time) or maximize road centrality (betweenness, closeness or information)? Or do they rather plan their routes according to the anchor-based model? By comparing computed optimal paths with the observed paths, especially when calculating the overlap between these two groups, it is possible to conclude if drivers try to minimize a certain attribute. Thus, research question (1) is used to assess if taxi drivers minimize travel distance or free-flow travel time. Question (2) is relevant, because it is important to examine what factors influence the behavior of taxi drivers, in what way they do so and how strong the effect is. The influence of weather conditions on route choice behavior is still unclear, thus the analysis of this third factor is especially interesting. The goal of question (3) is to see whether the anchor-based route choice model from Manley, Addison, and Cheng (2015) is also applicable to taxi drivers in Vienna. The analysis will reveal whether the anchor-based model is suitable for the explanation of taxi drivers in Vienna and how the potential differences in spatial knowledge between minicab drivers in London and licensed drivers in Vienna affect the results. Additionally, it will be investigated if there is some evidence for a better route choice of taxi drivers compared to routes chosen by ordinary drivers. Question (4) is important to see whether taxi drivers prefer some areas and avoid others.

In addition to these main research questions the results of the work are compared with the results of existing studies. It will be investigated if an analysis of a new and large-scale dataset confirms the findings of other studies.

3.2 Hypotheses

Before continuing with the description of the data, tools and methods used for the analysis, three hypotheses are listed that have been created before the analyses had been executed:

1. Taxi drivers do not follow shortest distance or shortest free-flow time paths.
2. Routes show distinct characteristics depending on time of the day and day of the week. Weather conditions do not have a significant effect on route characteristics.
3. Taxi drivers are very experienced and choose near-optimal paths, minimizing travel time. Thus, they plan their routes beforehand and only adjust them if there are unexpected deviations from the usual situation (e.g. a traffic jam due to an accident). Thus, directional effects are less clear than for the minicab drivers in London analyzed in Manley, Addison, and Cheng (2015).

At the end of the thesis, these hypotheses will appear again and will be discussed in relation to the achieved results.

4. Data and methodology

This section describes the methodology of this work in detail. In a first part, the data that has been used in the present thesis is described. Second, the tools that have been used to conduct this research are outlined. Third follows an in-depth description of the whole procedure, which has been followed to generate the results of this research. Last, some of the limitations of this work are discussed.

4.1 Data

There are three datasets used for the analysis in the study. First, we have the taxi trajectories, which are used to analyze the route choice behavior of cab drivers. Second, there is the street network data of Vienna, which has been used to map-match the trajectories. The network data is also used to generate shortest paths and to compare observed with computed paths. Third, there is the weather data, which is used in to evaluate the influence of the weather conditions on route choice behavior.

4.1.1 Floating car data

The core data consists of a database of more than two and a half million unique trips from taxi drivers in Vienna. These trips consist of 194’186’651 records in the database. The dataset was provided by the Austrian Institute of Technology, which in turn received the data from the private taxi company Taxi 31300 in Vienna. The trips were conducted during the months of June and November in 2015. The database consists of two tables called *lane measurement* and *trips*. The *trips*-table stores timestamps together with GPS positions, heading, covered distance and time difference to the previous record (see Table 4). On average, a GPS position is measured every 37 seconds during a trip. The *lane measurement*-table stores the routes reconstructed based on a map matching algorithm and contains derived routing and speed information (see Table 5). The map matching approach used to generate the present floating car data is described in Koller et al. (2015). The result of this process is the storing of a unique road-id for every segment of the trip, so that the complete trip can be traced on the OSM-street network. Importantly, the map-matching has been conducted beforehand and was not part of the present thesis.

Table 4: Data structure of the trips-table.

Field	Type	Comment
trip_id	Integer	Unique ID for each trip
timestamp	Timestamp	YYYY-MM-DD HH:MM:SS
longitude	Float	Longitude value (WGS 48 coordinates)
latitude	Float	Latitude value (WGS 48 coordinates)
heading	Integer	Heading at timestamp
covered_distance	Float	Distance covered since last timestamp
timediff_to_previous	Integer	Time passsed since last timestamp
geom	Geometry	Point geometry of lat/long-coordinate (WGS 84)

All trips in the database are such that have been executed with passengers on board. This has some very important implications as the route choice behavior of taxi drivers differs significantly depending on whether they are having passengers onboard (Sun, Karwan, and Kwon 2016).

At this point, some comments on the drivers and on the taxi business in Vienna in general are appropriate. To obtain a taxi license, drivers need to pass a test. Most importantly, the test contains a lot of questions about optimal routes from some location to another. Candidates also need to know the location of a lot of important landmarks (railway stations, embassies, hotels, police stations, media outlets, cafés, cemeteries and many more). Furthermore, potential taxi drivers need to know the name and location of many streets in Vienna.

A copy of an example test for a taxi-license in Vienna (in German) can be found under: https://www.wko.at/branchen/w/transport-verkehr/befoerderungsgewerbe-personenkraftwagen/Fragenkatalog-2018.pdf.

Unfortunately, data about the experience of the drivers is not available. Observations in Vienna have shown that the taxi drivers do usually not use in-car navigation systems.

Table 5: Structure of floating car data of Viennese taxi drivers (lane measurement-table; PK = Primary Key, FK = Foreign Key).

Field	Type	Comment
trip_id	Integer [PK]	Unique ID for each trip
timestamp	Timestamp [PK]	YYYY-MM-DD HH:MM:SS
route_element_index	Integer [PK]	Ordering of route segments in case of identical timestamp
road_id	Integer [FK]	id of the road segment
from_junction_id	Integer	osm id of the last visited node of the previous road segment
in_road_dir	Integer	boolean value indicating direction of travel
speed	Float	Derived speed at timestamp
length	Float	Length of road segment in degrees

It is also interesting to look at the distribution of pick-up and drop-off points. Figure 5 and Figure 6 show the density of drop-off and pick-up points respectively. Unsurprisingly, the highest densities are in the center of the city in both cases. There is one exception: The area around Schwechat airport in the southeast of the city contains a very high number of drop-off and pick-up points. Generally, once can see that the coverage south of the Danube is significantly better than it is north of the major river of the city.

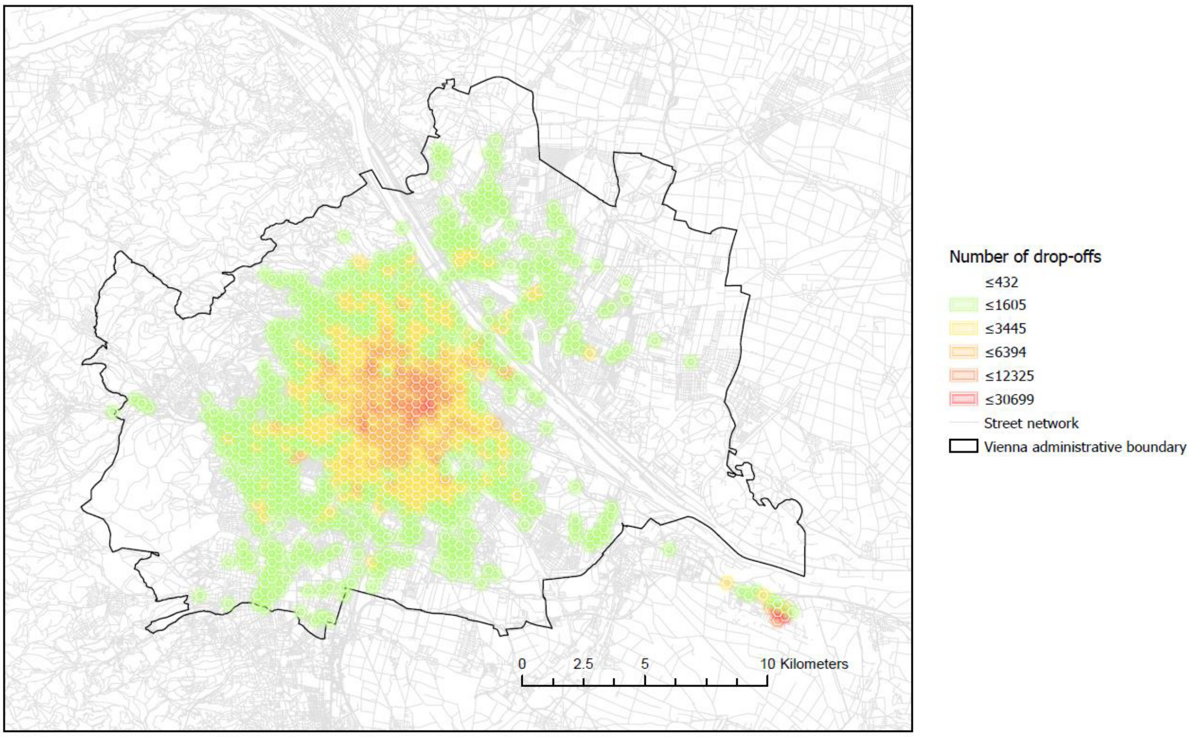

Figure 5: Hexagonal bins showing the density of drop-off points in Vienna.

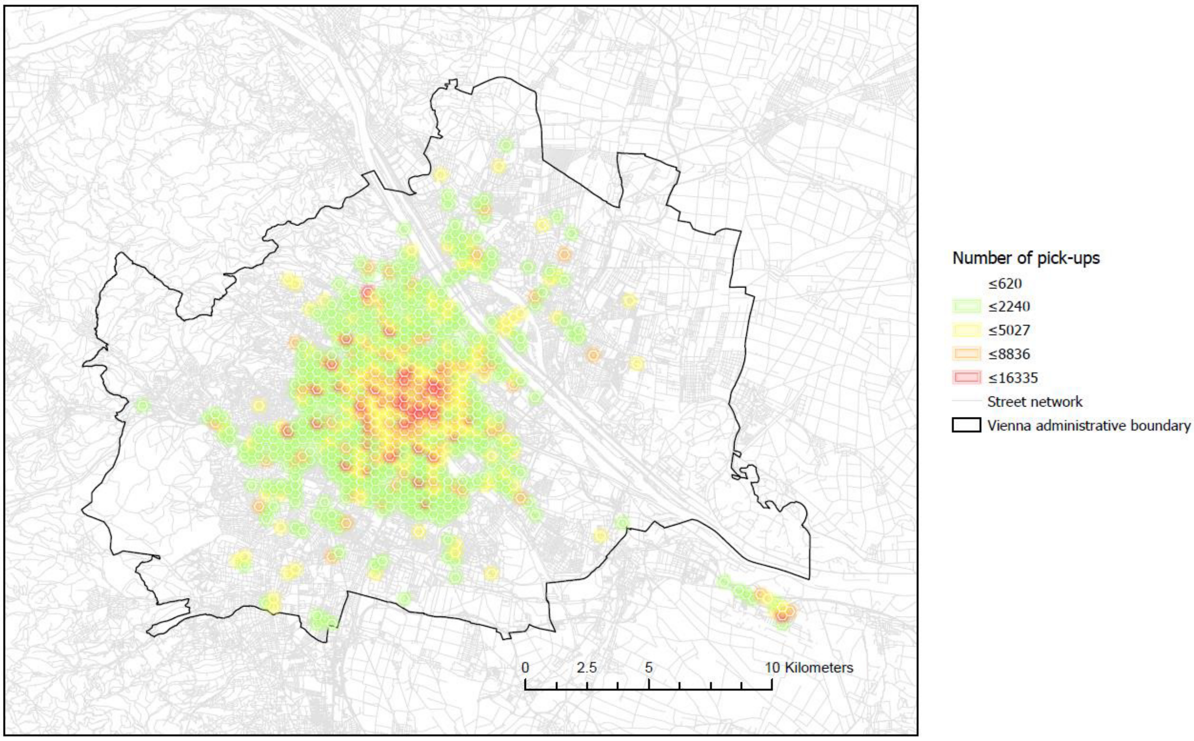

Figure 6: Hexagonal bins showing the density of pick-up points in Vienna.

4.1.2 Network data

The street network of Vienna extracted from OpenStreetMap (https://www.openstreetmap.org/) figures as the base map on which all the trips have been matched on. The network data has been extracted in 2015. Thus, the network data temporally matches the floating car data. Table 6 shows which attributes are stored in the OSM-street network data. The data includes information about one-way streets, road type and speed limits.

Table 6: Attributes of the street network data extracted from OSM.

Field	Type	Comment
gid	Integer [PK]	Unique id for gemetry of the road segment; Points to ways-table
id	Integer [FK]	Unique ID for road segment points to lm table
osmid	Integer [FK]	OSM ID of the road segment
fromjuncti	Integer	OSM ID of start node
tojunction	Integer	OSM ID of end node
name	Text	Street name
highway	Text	Road type of segment
geom	Geometry	Line geometry of the road segment

In addition to the core network data (road segments with associated information), turn restrictions from have been extracted from OpenStreetMap as well. The incorporation of these restrictions into the routing algorithm will be described further below. Unfortunately, these restrictions do not temporally match the rest of the data, as they have been extracted in late 2017. Therefore, there might be some restrictions in the dataset that were non-existent in 2015 and vice versa.

Figure 7: Primary, secondary and tertiary roads around Vienna.

The map in Figure 7 shows major highways in dark blue, secondary streets in orange and tertiary roads in light blue. There is a ring road around the historic center of the city, which is a one-way street and quite popular amongst taxi drivers. This ring road will appear a lot in the case studies in chapter 5.3.

4.1.3 Weather data

Information about weather conditions present in Vienna during the months of June and November 2015 has been downloaded from OpenWeatherMap (https://openweathermap.org/). On this webpage it is possible to get historical weather data for individual cities. The weather data includes information about precipitation, wind direction, wind strength and temperature. More importantly, the predominant weather conditions are stored on an hourly basis. Examples of such conditions are: “Clouds”, “Rain”, “Clear” and “Snow”.

4.2 Tools

For the analysis part of the thesis, mainly three tools have been used: PostgreSQL for storing, querying and analyzing the floating car, weather and street network data, Python for additional data manipulation and the computation of route characteristics and various GIS for analyses as well as visualizations.

4.2.1 PostgreSQL 10

The original data is already quite large (there are almost 200 million rows in the *trips*-table alone). Together with the paths that have been computed during the analysis, the data so large that it is necessary to handle data storage and analysis very efficiently. Therefore, the original trajectories as well as the computed shortest and fastest paths were stored, managed and queried in a database management system (DBMS). For this study, PostgreSQL, an open-source relational DBMS, has been used. PostgreSQL offers two extensions, which are essential for the present study: PostGIS and pgRouting.

The PostGIS extension adds a lot of functionalities and capabilities to store and query spatial data adequately to PostgreSQL. It adds new datatypes, such as geometry and geography, spatial indexes and query expressions based on location (Obe and Hsu 2015). The geography and the geometry datatype enable PostGIS to handle geographical coordinates. The geometry type is used for projected coordinates (e.g. ETRS89 in Austria, LV95 in Switzerland, or NAD83 for the US just to name a few), while the geography type is used for data with unprojected coordinates (only WGS84 is supported in the current version of PostGIS; see Obe and Hsu (2015). In other words, the geometry type is used for coordinates on a cartesian plane, the geography type is used for coordinates on an ellipsoid. The distinction is important as calculations of distances, areas and angles lead to different results depending on whether a cartesian or a projected coordinate system is used.

pgRouting is an extension to PostGIS and provides geospatial routing functionality to PostgreSQL databases. The pgRouting library contains various algorithms that can be used to generate all sorts of shortest paths. Most algorithms take an origin, a destination and some cost function as parameters and then, based on the cost, generate one or more shortest paths between the origin and destination. In this study the Turn Restriction Shortest Path (TRSP) algorithm was used to generate shortest distance and fastest travel time paths for all observed trips in the database.

4.2.2 Python

In this project, Python (version 2.7) has been used in various stages of the workflow. PyCharm has been used as an IDE (integrated development environment) and eight route characteristics for computed and observed trips have been calculated using Python. The psycopg-module was used to connect to the PostgreSQL database, load some data and calculate the characteristics. More importantly, the calculation of shortest distance and fastest travel time paths has been done with the use of a Python-script. Hereby, the multiprocessing module enabled the parallelization of the computations, which allowed for a up to four times faster calculations of the paths. The scipy-module was used to calculate some basic statistics when comparing characteristics of different groups of paths (standard deviations, means, t-statistics and p-values). Additionally, a lot of the data preprocessing work has also be done using Python and all the street network centrality indices were calculated using the NetworkX-module.

4.2.3 QGIS, ArcMap and ArcGIS Pro

All the visualizations in this study were created using Geographical Information Systems, namely QGIS, ArcMap and ArcGIS Pro. The former is open-source, the latter two are proprietary products. As every GIS of the three mentioned above has different strengths and weaknesses, always the one suited best for a specific task has been used. For example, computations of the local indicator of spatial association (LISA) have been made with ArcGIS Pro (see for example Figure 38). QGIS has been used a lot for initial explorations of street network data and computed and observed paths.

4.3 Study area

The floating car data spans over the city limits but most GPS records lie within the region displayed in Figure 8. All analyses in the case study section of this work are located within the city limits of Vienna with the area around Schwechat airport being the only exception.

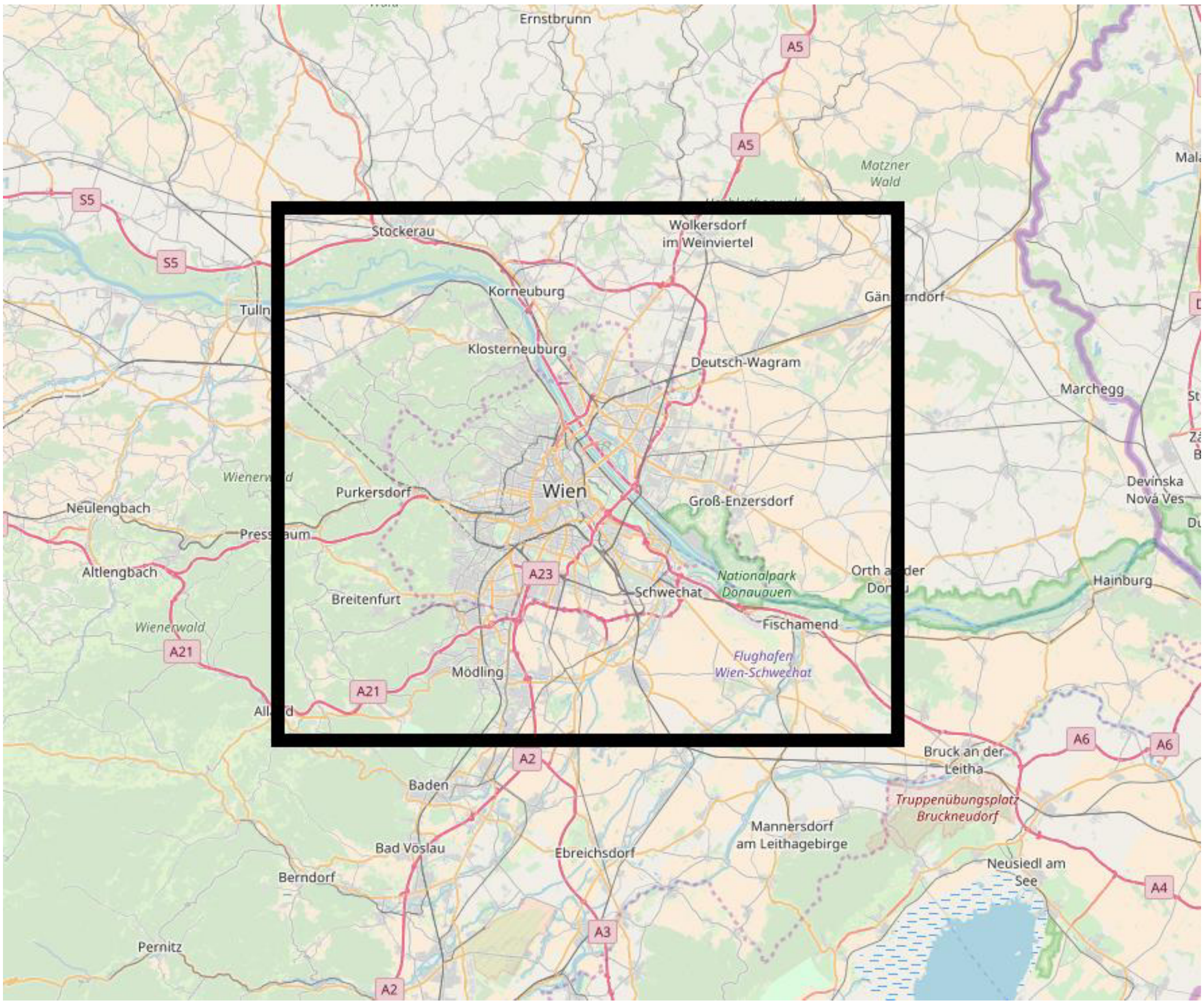

Figure 8: Analysis area (black box) encompassing Vienna and surrounding areas; map © OpenStreetMap contributors.

4.4 Procedure

Next, the procedure of this thesis is described. The aim of this is, that the conducted research is transparent and reproducible. First, the general workflow and the most important steps are discussed to give an understanding of the whole research from a high-level perspective. Second, a more detailed description of all steps is given. Figure 9 shows the whole workflow in a concise way.

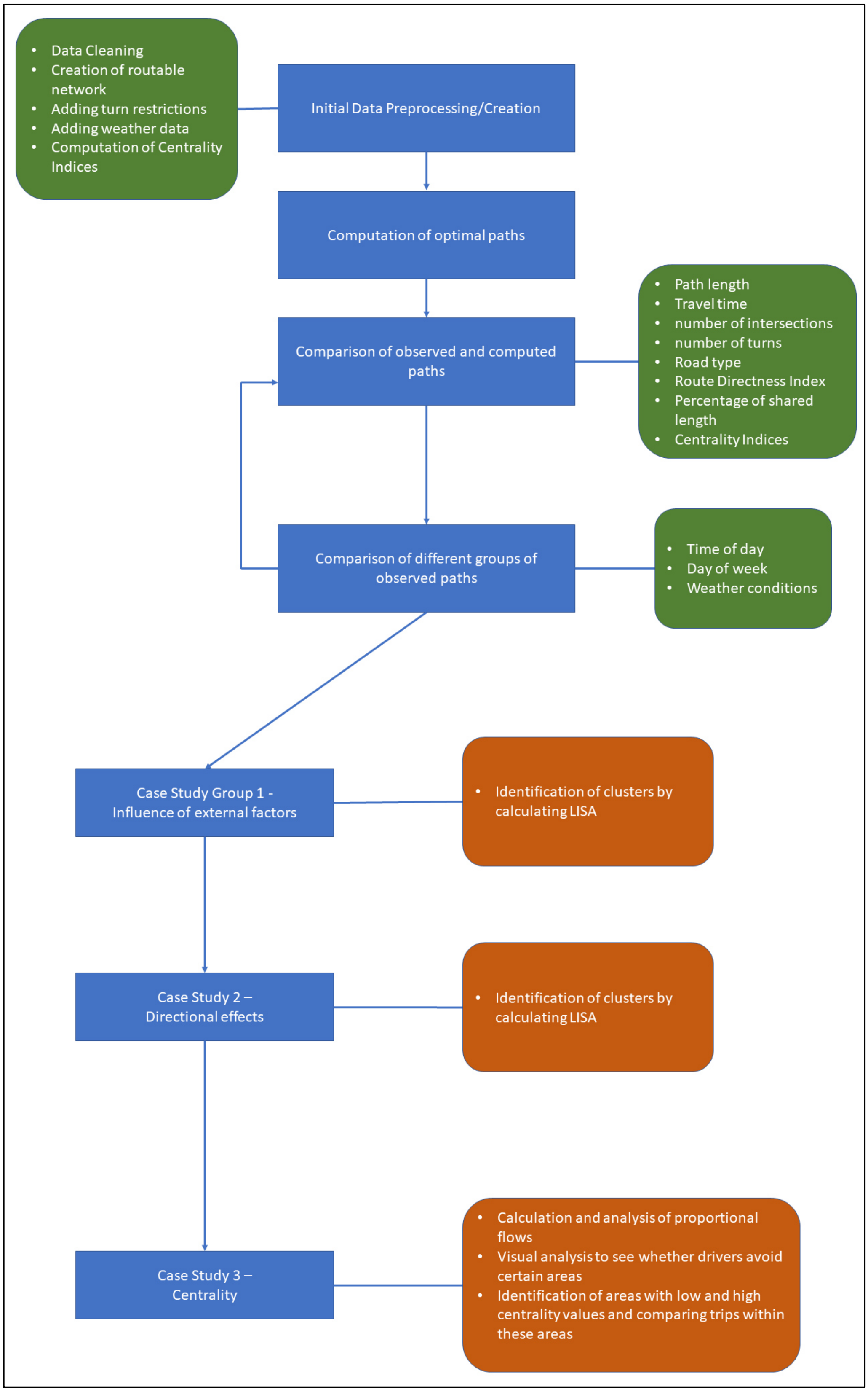

Figure 9: Schematic representation of the general workflow, which has been followed during the analysis stage.

First, the data at hand has been cleaned and additional data has been collected. Second, shortest distance and fastest free-flow travel time paths have been calculated. Third, the observed paths have been compared with the computed paths in terms of route characteristics. Fourth, all trips have been split into different groups and the previous step has been repeated for these groups. Last, the analysis has been finished with the examination of all case studies in the three groups.

4.4.1 Data preprocessing

Before starting the substantial part of the analysis, some data pre-processing and data creation was needed. The data pre-processing consisted of cleaning the network and floating car data. Streets that are not accessible to cars (e.g. bicycle lanes and footways) have been tagged so that they can be excluded from the analysis if necessary. For generating shortest distance and fastest travel time paths, it is necessary to identify start and end points for every trip as the routing algorithm needs these as an input. These have been extracted by selecting the entry with the “smallest” and “biggest” timestamp grouped by the trip IDs. The data creation part consisted of downloading the weather data and loading the resulting csv file into the database.

4.4.2 Creation of a routable network

Street network data has been loaded into the database using a tool called *osm2pgrouting*, which comes with the installation of the pgRouting-extension. This tool reads osm-files, which are created when exporting from OSM and stores the street network in the database. During this process, the tool automatically creates a routable network. It generates two cost attributes per road segment, one is based on the segment’s length and one is based on the segment’s speed limit. For each of these two cost attributes, there is also a reverse cost attribute with the same absolute value, however, it is negative in case of a one-way street. Edges with negative values are not considered by the various routing algorithms of the pgRouting extension. Thus, one-way street restrictions are added automatically by the osm2pgrouting-tool. Next, road segments which are not traversable by car (e.g. footway, bridleway, steps etc.) have been multiplied by -1 to exempt these segments from the routable network as well. Last, turn restrictions have been added by creating a separate restriction table. This table has the structure shown in Table 7. The first column shows edges which have restricted access from certain directions. The second column lists all edged from which it is not allowed to enter the edge from the first column. The value for the third column is 100’000 for every tuple in the table. It represents the cost to traverse this segment. As they are very high, the routing algorithm will always select routes not using invalid turns as paths around the restrictions always have lower costs.

Table 7: Format of the restrictions-table.

Field	**Type**	**Comment**
to_cost	Integer	Cost
target_id	Integer	To node
via_path	Text	List of from nodes

4.4.3 Calculation of centrality indices

Next, degree, closeness, betweenness and information centrality indices have been calculated for the street network of Vienna. Beforehand all streets not accessible to cars were removed from the graph representing the road network. First, the four indices are calculated for each node. There exist several ways to calculate edge centrality indices based on node centrality indices (Graser et al. 2016). In this study, edge centralities were calculated by averaging the values of the respective start and end nodes of an edge. Therefore, a primal representation for the street network was chosen in this work. Centrality indices can be calculated globally and locally, whereby each method has its own strengths and weaknesses. Information and betweenness centrality do not suffer from border effects very much.

However, closeness centrality does and thus for this measure a local computation is suggested in the literature (Porta, Crucitti, and Latora 2006; Graser et al. 2016). For this study, all indices have been computed globally.

4.4.4 Calculation of shortest paths

Shortest distance and fastest travel time paths have been calculated using the turn restriction shortest path (TRSP) algorithm. It is based on a simple Dijkstra-Shortest path algorithm. However, it also takes into account turns restrictions as the name already implies. The creation of the necessary table with the information about these restrictions was described in section 4.4.2 Creation of a routable network The length of a street segment has been chosen as the cost attribute for the calculation of the shortest paths, whereas a normalized speed limit attribute represents the cost for the calculation of the fastest free-flow travel time paths. The algorithm also considers the information from the restrictions-table which leads to invalid sequences of road segments having a very high cost, so that only valid paths are generated. As the optimal paths needed to be generated for all observed trips, the calculations need a lot of time. Therefore, the process was sped up by parallelizing the calculations. To do so, the multiprocessing module has been used. Normally, a Python process only runs on one CPU-core. With the multiprocessing module it is possible to go around this limitation by side-stepping the global interpreter lock (see https://docs.python.org/2/library/multiprocessing.html for more details) and spawn multiple Python processes allowing the use of all CPU-cores. As the machine that has been used for the computations has four cores, the computation of the paths could be speed up almost four times. This way, the amount of time needed to compute all paths could be brought down significantly. The task at hand was optimal for parallelization and an embarrassingly parallel computation was possible in this case. Table 8 contains the information that is stored for the resulting paths.

Table 8: Attributes of shortest and fastest paths stored in the database.

Attribute	Datatype	Comment
trip_id [PK]	Integer	Unique ID per trip
seq [PK]	Integer	Position of the segment within the path
cost	Float	Cost based on segment's length
cost_s	Float	Cost based on segment's speed limit
length_m	Float	Length of road segment in meters
one_way	Integer	0 = one-way, 1 = two-way street
class_id	Integer	ID for the road type of the segment
gid	Integer	Unique road segment ID
source	Integer	OSM-ID for the source node of the segment
target	Integer	OSM-ID for the target node of the segment
osm_id	Integer	OSM-ID for the segment
the_geom	Geometry	Geometry of the road segment

4.4.5 Computation of route attributes

To analyze chosen routes, compare them with optimal paths and to analyze the effect of external factors, route characteristics needed to be computed to be able to quantify the comparison of paths with each other. Thus, Python-scripts generating the following eight route characteristics have been created:

1. Trip distance
2. Driving time
3. Number of turns
4. Number of intersections
5. Road type
6. Centrality
7. Route Directness Index
8. Percentage of shared length

Trip distance was calculated by summing up the lengths of road segments for each trip. For the travel time characteristic, there are two distinct calculations leading to two results: Actual driving time for observed paths, theoretical time based on speed limits for observed and computed paths. Actual Driving time was computed by calculating the temporal difference between the last timestamp and the first timestamp of a trip. Theoretical driving times could be calculated by summing up the cost_s attribute of the roads segments for each trip, as this attribute represents the amount of time in seconds that it takes to drive along the segment assuming free-flow travel conditions.

The number of turns were computed by first calculating the angles between every road segment with the subsequent one starting from the first edge of a trip. If such an angle is bigger than 30 degrees, it is classified as a turn. If it the angle is more than 60 degrees, it is not only classified as a turn but also as a sharp turn. The orientation (left or right) of a turn can be determined by calculating the dot product between three points making up a turn. If it is bigger than zero, it is a left turn, if it is less than zero, it is a right turn (if it is equal to zero, then the three points lie on one line and there is no turn).

The number of intersections along a trip could be found out by counting how many nodes with an in-degree bigger than three a trip leads through. In other words, a node is classified as an intersection if there are at least four streets going out of it.

To get the relative share of primary, secondary and tertiary roads as percentages, all road types that existed in the database (which are tags extracted from OSM) have been put in one of these three categories. Then by summing up the length per road type and dividing it by the total length of a trip, the relative percentage per road type was calculated.

For each trip, the average degree, closeness, betweenness and information edge centrality were calculated. As the centrality values have been calculated beforehand (see chapter 4.4.3), these values could be extracted directly from the database.

Route directness has been calculated with the following formula:

$$RDI = \frac{total\ route\ distance}{straight\ line\ distance}$$

The total route distance can be computed by summing up the length of all road segments that are part of a trip and can be extracted directly from the database. The straight-line distance was calculated using PostGIS' st_distance-function which calculates distances on a spheroid when given WGS84 latitude/longitude-coordinates.

Percentage of shared length has been calculated with this formula:

$$PSL = \frac{length\ of\ shared\ links}{length\ of\ observed\ route}$$

4.4.6 Integration of external factors

For the next part of the analysis, all trips (observed and calculated paths) needed to be split into groups according to time of the day, day of the week, and weather conditions to assess the influence of these factors. This is done by again calculating the route characteristics described in the previous section for all groups of paths.

4.4.7 Comparison of groups of paths

For the comparison of route characteristics of the groups of paths, t-test are used to check how strong they differ and if the difference is statistically significant. More specifically, two-tailed paired sample t-tests have been used to compare observed with the corresponding computed paths. For the comparison of separate groups of observed paths (after they have been split according to external factors or by location), two-tailed independent t-tests have been used.

4.4.8 Case studies

The third and last step consists of a detailed in-depth analysis of the behavior of taxi drivers by using case studies. For this, interesting origin and destination areas have been identified and the trips between these regions were then analyzed in detail. These case studies help to reveal other factors that influence the behavior of taxi drivers, which cannot be captured by simply computing statistics over all trips. The case studies are put into one of the following three groups:

1. Influence of external factors
2. Directional effects
3. Centrality

Additionally, the overlap between observed routes for the same OD-pair has been explored beforehand. Three case studies are used to show how much the chosen routes of taxi drivers in Vienna overlap going from a specific origin to a specific destination. The higher the overlap, the more homogenous the drivers are and vice versa. If the overlap is high this would mean that the drivers are on a similar level in terms of experience and spatial knowledge. This is not classified as a separate case study group, as the influence of spatial knowledge is not a strong focus of this thesis and no information about driver's knowledge is available for the FCD. However, it is important to briefly check the variability in chosen routes to put other results into perspective.

Case Study Group 1 – Directional effects

In this first group, directional effects are looked at. Manley, Addison, and Cheng (2015) found that minicab drivers in London often follow different paths going from a point A to a point B as opposed to going from B to A. As this is not a result of constraints of the street network (e.g. one-way streets, turn restrictions), they deduct that this is a sign of anchor-based route choice. The idea behind this framework is that people do not necessarily plan their routes beforehand, but that they plan their routes on-the-fly while driving (see summary of the framework in chapter 2). More importantly, people usually do not seem to follow any minimum cost paths. Thus, such theoretically optimal paths perform badly when trying to predict the route choice behavior of drivers. According to the proposed framework, people rather plan their route around anchors - which usually are important intersections - where multiple streets meet, and the drivers must choose from many options when continuing their journey. At these anchors the drivers identify the best route to the next anchor, drive there and repeat the process until they arrive at the destination.

It will be investigated if there are such directional effects present in this thesis' dataset. There are two potential differences between the data that Manley, Addison, and Cheng (2015) analyzed and the data from Vienna. First, the spatial knowledge and the experience between taxi drivers in Vienna and the minicab drivers in London might differ. Second, the layout of the street network of Vienna is different from the one in London. These two points could influence the route choice behavior and leads to the following hypotheses.

Hypotheses

As the taxi drivers in Vienna are better trained and more experienced than the minicab drivers, they might not follow the anchor-based route choice framework. Therefore, directional effects are expected to be less clear than in the Manley, Addison, and Cheng (2015) study. There might be scenarios where directional effects are clearly visible, but this might not be the case for other scenarios.

Methodology

For the scenarios in this group, the same approach as in Manley, Addison, and Cheng (2015) is followed: Road segment based calculation of differences (trips going from A to B minus trips going from B to A), and area-based identification of clusters with high mono-directional traffic. These clusters can be found using LISA.

The origin and destination areas are determined with the use of an SQL-query. It selects *n* nearest points around an origin coordinate as well as *m* nearest points around a destination coordinate. Then, IDs from all trips whose start-point lies in the first set and whose end-point lies in the second set, are selected. Thus, by changing the values for *n* and *m,* it is possible to adjust the size of the resulting set of trips. However, the bigger *n* and *m,* the bigger get the origin and destination areas. Very big areas should be avoided as all trip origins and destinations should be close together, so that the choice set is the same for all trips.

Case Study Group 2 – Influence of external factors

The focus of the second group of case studies is the influence of the following external factors: Time (time of the day and day of the week) and weather conditions. Comparing just the route characteristics from the trips grouped according to the above categories, reveals some differences in the case of time of the day and none for the weather conditions. However, there might still be differences which remain hidden until one looks at where the chosen paths actually pass through. When two paths exhibit similar route characteristics that does not necessarily mean that these two paths also highly overlap. Thus, taxi trajectories for the same OD-pair may be completely different between for example rush-hour and nighttime (Yang et al. 2017).

Hypotheses

Drivers in Vienna take different routes from A to B depending on time of the day and day of the week, most likely because the amount of traffic varies. Therefore, route choice behavior is influenced by some external factors. The influence of weather conditions is probably not visible in the analysis.

Methodology

The same methodology as in case study group 1 is applied here. The differences will be calculated per road segment as trips from group 1 minus trips from group 2 (for example trips during daytime minus trips during nighttime going from A to B). Again, clusters can then be identified using LISA calculations.

Case Study Group 3 – Location, centrality and street network configuration

Last, multiple centrality indices are incorporated into the analysis (betweenness, information and closeness centrality). In this group, in addition to varying the location of origin and destination, the investigated centrality index changes as well between different scenarios.

Specifically, the following shall be quantified using centrality indices:

- Areas that drivers tend to avoid (e.g. drivers avoid areas with high/low information or closeness centrality)
- Streets that are popular amongst taxi drivers. More specifically, there may be some correlation between popular routes and high betweenness centrality values.

Additionally, it is investigated if the chosen routes exhibit different characteristics depending on the region there in (areas with similar information/closeness centrality values – clusters of low or high values – form such regions).

Hypotheses

- When going from A to B, most drivers tend to avoid certain regions X between A and B. These regions X exhibit relatively lower information/closeness centrality values than the surrounding areas.
- Taxi drivers prefer streets with high betweenness centrality values.
- Trips in areas with high information/closeness centrality differ significantly from trips in areas with low information/closeness centrality in terms of route characteristics.

Methodology

Here, the methodology differs according of the focus of the scenario.

Focus: Areas that are avoided

1. Selection of an area A and an area B, such that a region X with relatively low information/closeness centrality lies in between A and B.
2. Visually check if drivers avoid area X.

Focus: Correlation with betweenness centrality

1. Extract traffic flows from an area A to an area B.
2. Compare these flows with the road network classified by betweenness centrality.
3. Check if there are any correlations between high traffic flows and high betweenness centralities on a road segment-basis by visual exploration.

Focus: Different route characteristics in different areas

1. Identify a region A with high information/closeness centrality and a region B with low information/closeness centrality.
2. Calculate route characteristics for trips which entirely lie within area A or B respectively.
3. Compare route characteristics from trips in region A with trips from region B.

4.5 Limitations

This work as some important limitations, which need to be kept in mind when assessing the validity of the results in this thesis. First, it is difficult to relate the results to all traffic participants, as they might only be valid for taxi drivers and do not reflect the route choice behavior of ordinary drivers. Second, all drivers are treated as a homogenous group, as the trips are not assigned to any drivers in the dataset. In other words, the influence of inter-personal differences on route choice behavior is not part of the analysis in this thesis. Thus, differences in the level of spatial knowledge, personal preferences and characteristics cannot be accounted for. Third, there is no live traffic and no historical traffic data is incorporated. Construction sites and accidents can have a huge influence on route choice behavior, but these effects were not analyzed in this study. Last, a case study approach cannot proof that any hypothesis is true, it can only falsify them as long as not every possible case study (meaning every possible combination of origin and destination areas) of is looked at.

5. Results

In this chapter, the results from the research outlined in the previous chapter are displayed. In the first stage, the results of the comparison of the route characteristics of the observed and computed paths are shown. In the second stage follow the results of the analysis of the influence of the previously defined external factors on these route characteristics. In the last stage, the results for the case studies are presented.

5.1 Comparison of observed paths and computed paths

Table 11 shows the resulting t-values and confidence levels for the differences between the groups. Route characteristics have been computed for 20'000 randomly selected observed and the corresponding computed paths (shortest distance paths and minimum free-flow time paths). Dependent t-tests were used to check whether the differences between these groups are statistically relevant. Table 9 shows the values for the route characteristics of the observed and computed paths. The results for each route characteristic will now be discussed in detail.

Table 9: Values for variables of observed, shortest and fastest paths.

Variables	Observed paths (mean ± std)	Shortest paths (mean ± std)	Fastest paths (mean ± std)
Time (min) (free flow time)	6.81 ± 5.93	5.53 ± 5.65	4.98 ± 4.72
Time (min) (Actual travel time)	11.73 ± 11.05	NA	NA
Distance (m)	5655 ± 6087	4244 ± 4886	4432 ± 5230
Number of intersections	3.48 ± 3.88	4.02 ± 5.60	3.46 ± 3.86
Route directness index	1.59 ± 0.85	1.25 ± 0.23	1.30 ± 0.25
Turns statistics			
Left turns	1.18 ± 1.90	1.09 ± 1.61	1.12 ± 1.59
Right turns	1.12 ± 1.81	1.13 ± 1.66	1.23 ± 1.72
Total normal turns	2.30 ± 2.83	2.22 ± 2.92	2.35 ± 2.88
Sharp left turns	1.76 ± 2.06	2.63 ± 2.49	2.09 ± 1.92
Sharp right turns	1.82 ± 2.09	2.35 ± 2.25	1.85 ± 1.75
Total sharp turns	3.58 ± 3.42	4.98 ± 4.29	3.94 ± 3.14
Total all turns	5.88 ± 5.47	7.20 ± 6.30	6.29 ± 5.18
% of route based on road type			
% distance on primary roads	0.34 ± 0.35	0.25 ± 0.31	0.31 ± 0.34
% distance on secondary roads	0.26 ± 0.30	0.28 ± 0.31	0.30 ± 0.32
% distance on tertiary roads	0.40 ± 0.35	0.47 ± 0.35	0.39 ± 0.35
Overlap with corresponding optimal paths			
% of shared length with shortest paths	0.52 ± 0.38	NA	NA
% of shared length with fastest paths	0.53 ± 0.39	NA	NA
Centrality			
Betweenness centrality	0.00321 ± 0.00604	0.00205 ± 0.00371	0.00293 ± 0.00639
Closeness centrality	0.00829 ± 0.00047	0.00944 ± 0.00050	0.00948 ± 0.00050
Information centrality	1.0551-06 ± 8.4717-08	1.2193-06 ± 8.1628-08	1.2178-06 ± 8.4782-08
Degree centrality	2.9595-05 ± 2.6280-06	3.0246-05 ± 1.8384-05	2.9476-05 ± 1.7263-05

Route length

Unsurprisingly, the shortest distance paths are indeed shorter than both observed and fastest free-flow travel time paths. The length of the fastest paths is on average much closer to the length of the shortest paths than to the length of the observed paths. The latter are 33 percent longer on average than the shortest paths and 28 percent longer than the fastest paths. The difference between the two groups of computed is only four percent. The fact that the observed paths are much longer than the shortest distance and fastest free-flow travel time paths suggests that people do usually not follow these paths and that there are other factors which influence the route choice behavior. Furthermore, taxi drivers do not seem to minimize travel distance, which confirms similar findings from other studies. However, compared to the results of these other studies, the deviations are bigger. Papinski and Scott (2011) found the observed paths to be 13 percent longer than the shortest distance paths and ten percent longer than the fastest travel time paths. However, in that study, the routes are on average more than three times longer than is the case for the present dataset from Vienna's taxi drivers. Ciscal-Terry et al. (2016) reports the observed paths to be 19 percent longer than the matching shortest paths and 12 percent longer than the fastest paths. Again, the trips are significantly longer than the ones in the present study.

Time

Since real travel times are naturally only available for observed paths, the analysis of the results is confined here to the theoretical free-flow time values (real travel times are useful for the comparison of the distinct groups of observed trips which is discussed later).

The differences between shortest, fastest and observed paths show a similar pattern as before. The difference is that now it is the fastest paths which have the lowest values. Again, the mean values of the computed fastest and shortest paths are quite close to each other (11 percent difference), while the observed trips theoretically took 23 percent (shortest paths) or 37 percent (fastest paths) more time to complete. Since the streets are almost never empty and there is always some traffic, free-flow conditions are rare and these differences to the computed paths are not surprising. As we will see later during the second stage of the analysis, free-flow travel times and actual travel times usually show different patterns and do not correlate. Figure 10 shows the actual travel time of observed trips and hypothetical travel times for computed paths.

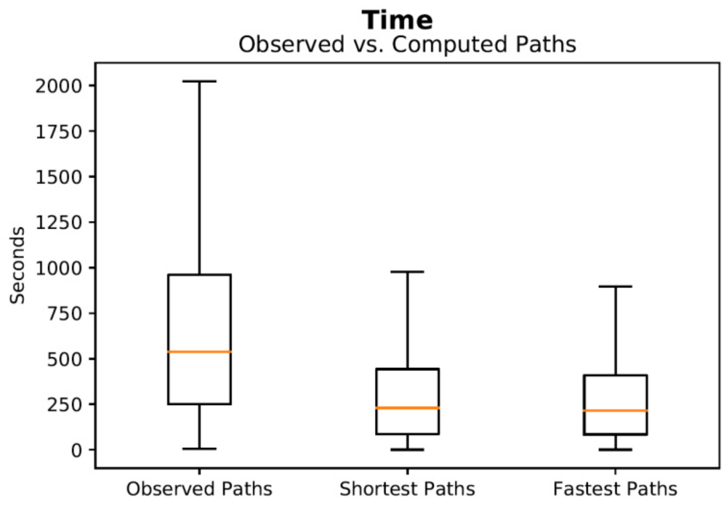

Figure 10: Values for travel time per group.

Number of intersections

For this variable, the values for observed (3.48) and fastest paths (3.46) are close together, while the shortest paths (4.02) have significantly more intersections along the route. This could be due to shortest paths having bigger shares of tertiary roads than observed and fastest paths (see results for the road type variable further below). Since the density of the street network of such smaller roads is usually higher than the density of higher order roads, the shortest paths – relatively using more tertiary roads – can be expected to have more intersections along them. Contrarily, the spacing between intersections is usually bigger for highways than it is for local roads. Thus, it makes sense that the fastest and observed paths, which lead relatively more often along such higher order roads, have fewer intersections in comparison.

Figure 11 shows a scenario where the shortest distance path goes through an area with a fine grid of tertiary roads while the observed path leads along the higher order road which goes around the area (the theoretically fastest paths in this scenario highly overlaps with the observed path and is left out for better legibility). When counting the intersections, it can be confirmed that the shortest distance path goes through more intersections than the observed path (in this study, a node is classified as an intersection when it has an in-degree of four or bigger).

Figure 11: Example of an observed trip with fewer turns than the associated shortest distance path.

Route directness index

The RDI values show a similar pattern as the distance values. While the shortest paths are the most direct (RDI: 1.25) and observed paths the least direct (1.59), the fastest paths are almost as direct as the shortest paths (1.30). This leads to the conclusion that taxi drivers do not necessarily try to find the most direct way to the destination. The results are very close to the numbers reported for non-professional drivers in Ciscal-Terry et al. (2016) and to those in Papinski and Scott (2011). Hence, driving experience does not seem to have a significant effect on resulting RDI values.

Figure 13 illustrates the differences between the three groups of paths regarding the RDI values. As observed and fastest routes tend to follow higher order roads that allow for higher speeds, these are generally less direct than the shortest paths, which by definition, represent the most direct routes.

An example helps to understand the differences in RDI values (see Figure 13). Starting at origin area *A* most taxi drivers first follow the southern road (*C1*), then turn left and then follow the street more or less straight to destination area *B*. In this scenario taxi drivers seem to minimize the amount of turns at the cost of more direct routes. By contrast, depending on the exact starting point, shortest distance paths either follow the street below *C2* involving a lot of turns, or the straight street above *C2*. Both these street sectors are also present in the observed paths, but they are not as popular as the southern street *C1* mentioned before. The theoretically fastest paths almost completely follow the straight street *C3.* A lot of these fastest paths involve detours (first increasing the distance to the destination, so that they can access the road *C3,* which allows for higher speeds).

Turns

In terms of the total number of turns irrespective of the direction and sharpness of the turns, shortest paths exhibit the most turns along their routes (7.2), followed by fastest (6.29) and observed paths (5.88). This is unsurprising since the most direct path in a street network usually involves a lot of turns and leads to routes with a lot of zig-zagging (see Figure 11 for an example). In contrast to time and distance, the fastest paths are more similar in terms of the number of turns to the observed paths than to the shortest paths. The number of sharp turns relative to the number of all turns is significantly higher for shortest paths than for the other groups. Again, this might be due to direct routes involving a lot of zig-zagging, which in the case of a regular shaped street network involving a lot of 90-degree angles, means that more turns are classified as sharp (a turn is defined as sharp when it has a turning angle greater than 60 degrees). It is due to these sharp turns that the shortest paths have more turns in total, the number of normal turns (angle between 30 and 60 degrees) is almost the same for all groups of paths. Regarding the direction of the turns, it is difficult to see any distinct patterns. The difference between the number of (sharp) left turns and the number of (sharp) right turns is small for all groups. There are slightly more left than right turns for the computed paths while there are exactly as many right turns as there are left turns on observed routes. This gap between computed and observed paths could mean that people prefer right over left turns since left turns involve the observation of the traffic in two directions while right turns involve only one direction that drivers have to pay attention to (of course this is only true for areas with right-hand traffic). Venigalla, Zhou, and Zhu (2017) have also found the computed paths to have more left turns than the observed paths. Figure 12 shows the results visualized with boxplots.

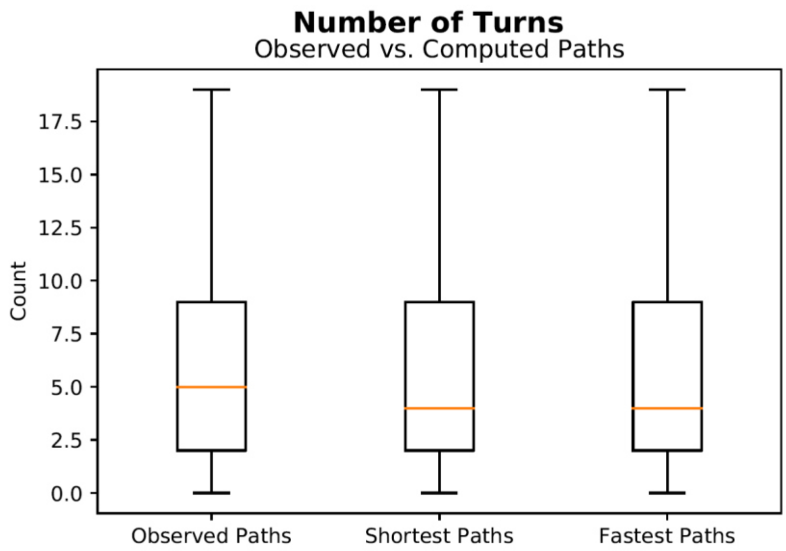

Figure 12: Results for turns per group.

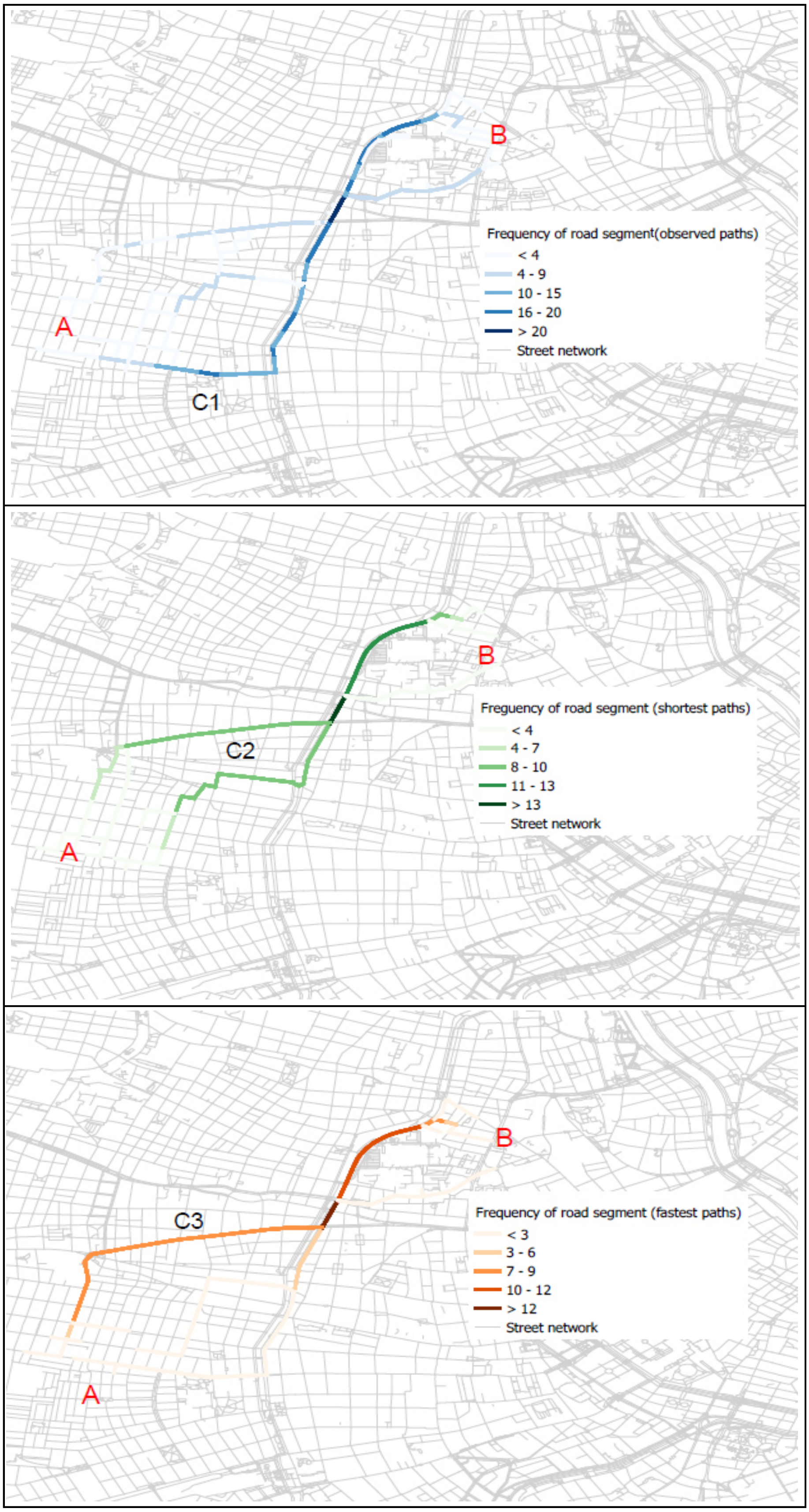

Figure 13: Comparison of directness of routes going from A to B.

Road type

The main difference between the groups concern the use of primary and tertiary roads. Shortest paths use a significantly higher portion of tertiary road segments than the other groups. At the same time, they make the least use of primary roads. Primary roads are more common in observed paths than for any of the groups of computed paths. However, primary roads are never the predominant category, the share of tertiary roads is the highest in all groups of paths. This is in line with findings of several other studies. Venigalla, Zhou, and Zhu (2017) report the use of tertiary (local) roads to be the highest (compared to the use of primary and secondary roads) for all groups of paths, except for very long paths (paths longer than 10 miles). The use of primary and secondary roads gets naturally higher when driving from town to town, since such trips usually involve the use of highways or other higher order roads that connect villages, towns and cities. In an intra-urban setting, like that in Vienna, this is not the case and as the data shows, taxi drivers prefer local roads over both secondary and primary roads. In Figure 14, Figure 15 and Figure 16 the results are shown as boxplots (the brown line indicates the median).

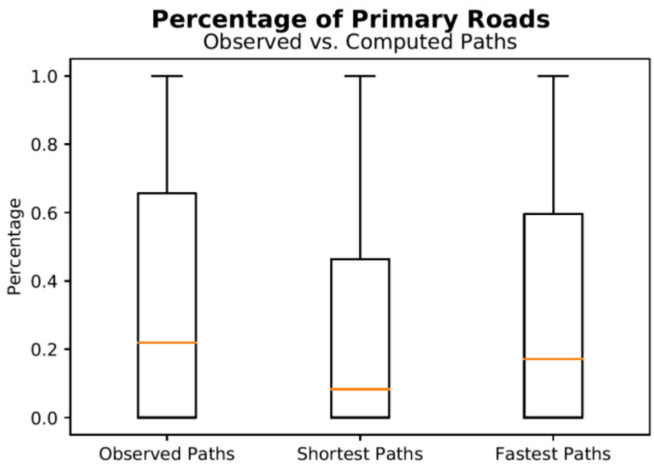

Figure 14: Primary roads per group.

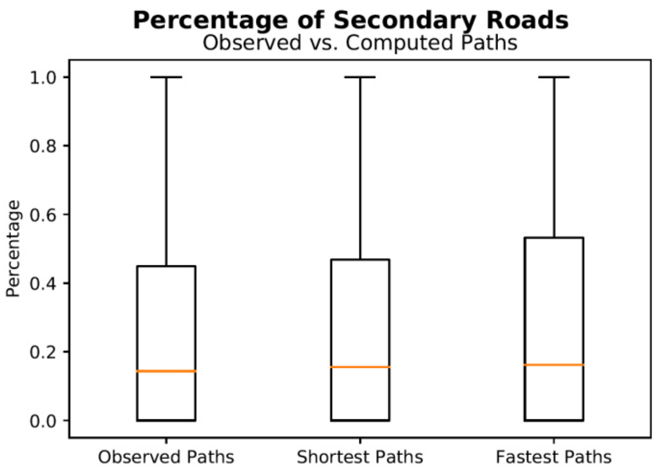

Figure 15: Secondary roads per group.

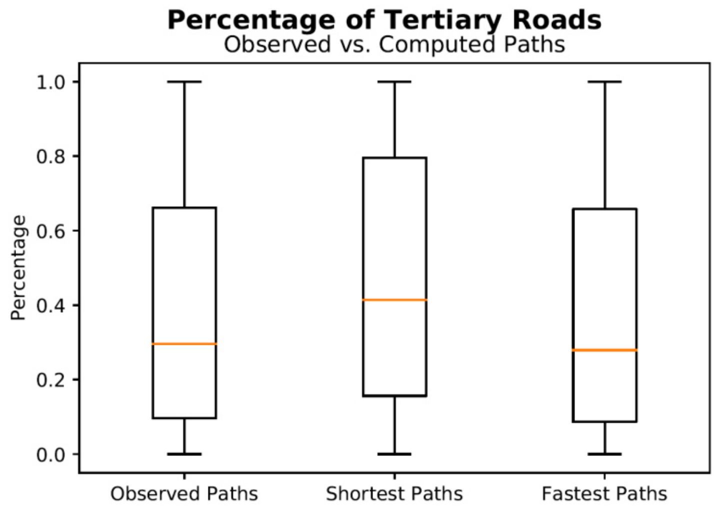

Figure 16: Tertiary roads per group.

Percentage of shared length

The percentage of shared length between observed and computed paths is just over 50 percent in both cases. This is in line with the results of other studies, which report low overlap between optimal distance and travel time paths (Sun et al. 2014; Zhu and Levinson 2015; Manley, Addison, and Cheng 2015). What is a bit surprising is that observed paths overlap almost as much with shortest paths than they do with fastest paths. For most variables the difference between the fastest and the observed paths is smaller than the difference between the shortest and the observed paths (this is true for all characteristics except free-flow travel time, and for closeness centrality. Considering this, the PSL values could be expected to be significantly higher for the fastest paths than for the shortest paths. However, this is not the case. This means that while the fastest and observed paths show (relatively) similar route characteristics, they do not follow along the same road segments. Figure 17 shows a scenario where all groups of paths are very similar, Figure 18 shows a scenario where the PSL-values are rather low.

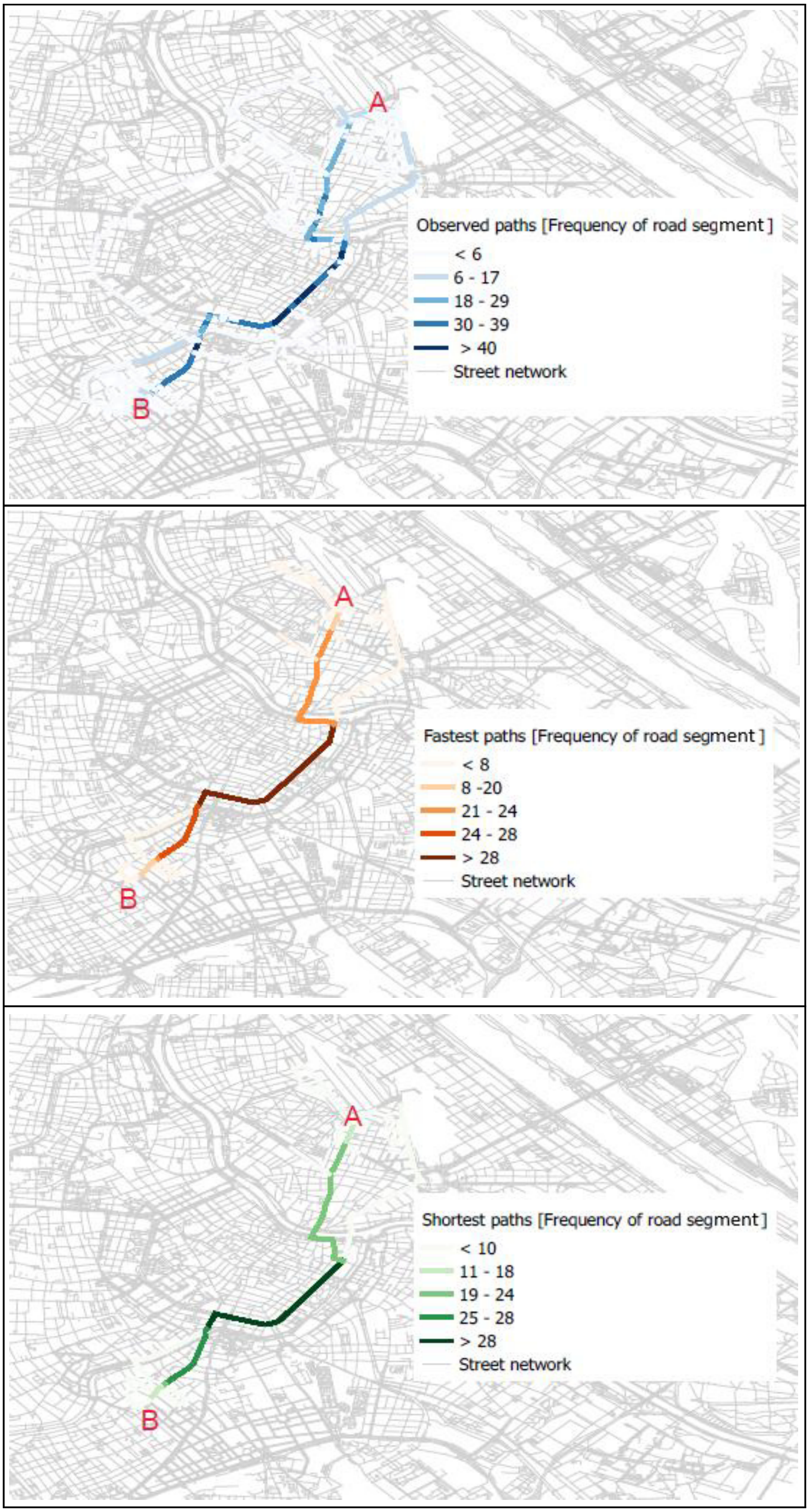

Figure 17: Scenario with high overlap between observed and computed paths going from A to B.

However, this is of course dependent on the location of origin and destination. There are scenarios with high overlap between computed and observed paths and there are scenarios without such overlap and lastly there are also scenarios, where only one of the two computed paths overlap with the observed ones. Figure 17 and Figure 18 show two such scenarios. In the first one, observed and computed paths highly overlap. In contrast there is almost no overlap in the second scenario. The influence of space on these differences and on the behavior of taxi drivers is part of the third stage of the analysis. However, it makes sense to look briefly look at two example scenarios, to illustrate what the results look like in space.

Figure 17 shows a case where there are few alternatives and thus, the PSL is high for all groups. Here, the origin area *A* lies north of the city center, the destination *B* is located south of the center. As sizeable parts of Vienna's city center are car-free, taxi drivers must drive around it. In this case, the relative location of origin and destination defines whether they drive around the center clockwise or anti-clockwise. In the scenario depicted here, they drive around it in clockwise direction, since this minimizes the travel distance. Based on the constraints in the dataset (turn restrictions, one-way streets and car-free zones) the shortest as well as the observed paths highly overlap. Additionally, the minimum free-flow travel time paths are equal to the minimum travel distance paths. Therefore, all three groups of paths highly overlap in this scenario.

Contrarily, the scenario depicted in Figure 18 shows the most popular route per group to be distinct from each other. Almost all taxi drivers first stay on the side of the Danube where the trip started, then follow the river on the highway A22 and in the end, they cross it using the Reichsbrücke. The fastest paths show a similar start, but they cross the Danube one bridge earlier, going over the Brigittenauer Brücke. Looking at the fastest paths in Figure 18, one can see that they are more direct than the observed paths. Since the distance is shorter compared to the observed paths and the speed limits are the same, these paths are a little bit faster in theory. However, this is either not the case, or the time gain is not worth the additional complexity (the fastest paths involve a lot more turns than the observed paths).

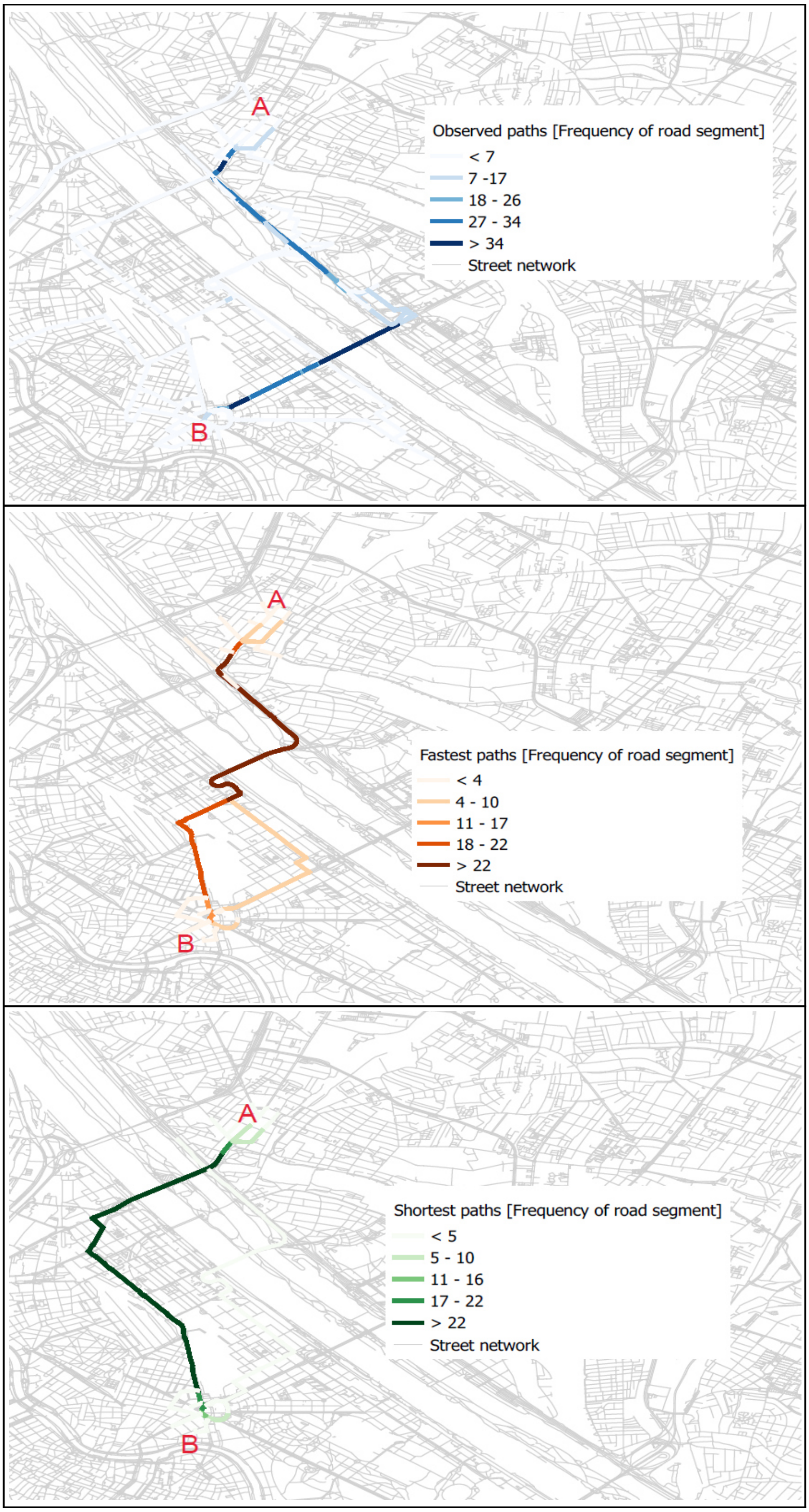

Figure 18: Scenario with low overlap between observed and computed paths going from A to B.

Centralities

Since centrality is a somewhat abstract concept, is not as straightforward as for the other variables to interpret the results. Generally, the centrality values per group should correlate so some extent with the percentage of road types used. For example, the more primary roads a group of trips uses, the higher should be the average betweenness centrality value of these paths. Comparing the average centrality indices per road type in Table 10 with the centrality values in Table 9 shows that this assumption is at least partially true. The betweenness values correlate with the road types used. For the other centrality indices, it is difficult to judge, as these do not differ very much between road types. In this light, it is interesting to see that compared to computed paths, observed trips have significantly lower average information and closeness centrality values. In terms of degree centrality, shortest paths have the highest value, which leads to the assumption that tertiary roads have a higher number of intersections per kilometer than secondary and primary roads.

Table 10: Average centrality indices per road type.

	Road Type		
Average Centrality	Primary	Secondary	Tertiary
Degree Centrality	0.000028836496	0.000029323302	0.000028985414
Closeness Centrality	0.007705139414	0.007536209125	0.007451438219
Betweenness Centrality	0.003405943432	0.002081071116	0.001361268929
Information Centrality	0.000000930905	0.000000938293	0.000000920019

General observations and summary

As already stated before, the values of the observed paths are closer to the values fastest paths than the shortest paths for the majority of the characteristics (namely distance, road type, number of intersections, RDI, number of turns and information centrality). Only for the travel time and the remaining three centrality indices the values of the shortest paths are closer.

Almost all the differences are statistically significant (between the observes and computes paths as well as between the two groups of computed paths; see

Table 11). Exceptions are the number of intersections between observed and fastest paths, the number of right turns with a turning angle less than 30 degrees between observed and shortest paths, the number of sharp right turns between observed and fastest paths, the average information centrality between observed and fastest paths and the average degree centrality between observed and shortest paths.

The hypothesis regarding the tendency of drivers to avoid left turns mentioned in chapter 3, can be confirmed as it seems to be true. The results show that indeed relative to shortest and fastest paths, observed paths contain fewer left turns.

To sum up the first stage of the analysis, we can conclude that in terms of route characteristics, there are significant differences between fastest, shortest and observed paths. This confirms findings from many studies that have shown these computed paths to be bad predictors of observed traffic behavior. Looking at it a bit closer, the fastest paths share more similarities with the observed paths than the shortest paths do. However, as the discussion of the PSL characteristics has shown, they do not share a lot of roads segments (at least not more than they do with shortest paths). Therefore, there must significant spatial differences. These will be investigated in the third stage of the analysis.

Table 11: t-test statistics and corresponding p-values for the differences between observed and computed paths. Bold values indicate differences significant at the 0.05 significance level (n = 20'000).

Variables	Observed vs. shortest paths	Observed vs. fastest paths	Shortest vs. fastest paths
Time (free flow)	**67.48 (0.000)**	**101.52 (0.000)**	**59.16 (0.000)**
Time (actual travel time)	NA	NA	NA
Distance	**86.11 (0.000)**	**78.01 (0.000)**	**-52.26 (0.000)**
Number of intersections	**-17.49 (0.000)**	0.85 (0.363)	**20.86 (0.000)**
Route directness index	**-60.96 (0.000)**	**-53.01 (0.000)**	**72.54 (0.000)**
Turns statistics			
Left turns	**7.88 (0.000)**	**4.95 (0.000)**	**-4.29 (0.000)**
Right turns	-1.34 (0.180)	**-8.63 (0.000)**	**-12.9 (0.000)**
Total normal turns	**4.72 (0.000)**	**-3.07 (0.002)**	**-10.4 (0.000)**
Sharp left turns	**-48.97 (0.000)**	**-21.63 (0.000)**	**43.68 (0.000)**
Sharp right turns	**-30.93 (0.000)**	-1.75 (0.081)	**45.08 (0.000)**
Total sharp turns	**-51.81 (0.000)**	**-16.17 (0.000)**	**52.56 (0.000)**
Total turns	**-37.52 (0.000)**	**-13.39 (0.000)**	**38.49 (0.000)**
% of route based on road type			
% distance on primary roads	**52.81 (0.000)**	**17.94 (0.000)**	**-49.32 (0.000)**
% distance on secondary roads	**-9.55 (0.000)**	**-22.5 (0.000)**	**-16.92 (0.000)**
% distance on tertiary roads	**-40.00 (0.000)**	**5.11 (0.000)**	**58.99 (0.000)**
Overlap with corresponding optimal paths			
% of shared length with shortest paths	NA	NA	**-5.01 (0.000)**
% of shared length with fastest paths	NA	NA	**-5.01 (0.000)**
Centrality			
Betweenness centrality	**31.98 (0.000)**	**8.50 (0.000)**	**-25.85 (0.000)**
Closeness centrality	**-617.86 (0.000)**	**-703.58 (0.000)**	**-35.02 (0.000)**
Information centrality	**-467.38 (0.000)**	**-445.65 (0.000)**	**10.98 (0.000)**
Degree centrality	**-4.95 (0.000)**	0.96 (0.337)	**12.21 (0.000)**

5.2 Influence of external factors

Until now we have only looked at the characteristics of observed paths in comparison to the corresponding computed paths irrespective of external factors. By grouping the observed paths according to weather conditions present during the trip, time of the day and day of the week, the influence of these external factors can be analyzed. For each factor it will be briefly described how the paths have been grouped before the results of every group of paths are characterized. The number of trips per group is displayed in Table 12. Since a trip can be in more than one group (for example a trip starts during good weather conditions and during the trip it starts to rain), the total number of paths per external factor is not always the same.

Table 12: Number of observed trips per group.

Group	Number of trips
Bad weather	380'903
Good weather	1'734'218
Rushhour	699'494
Daytime	1'177'516
Nighttime	822'134
Weekdays	1'806'598
Weekends	785'837

5.2.1 Time of the day

Three groups have been created for time of day: Rush-hour (06:00 to 09:00 and 16:00 to 19:00), nighttime (22:00 to 06:00) and daytime (09:00 to 16:00 and 19:00 to 22:00). Trips have been assigned to multiple groups if they fall into multiple time periods. There are 1’177’516 trips in the daytime group, 822’134 trips in the nighttime group and 699’494 trips in the rush-hour group.

Table 13: Route characteristics for nighttime, daytime and rush-hour trips.

Variables	Night trips (mean ± std)	Day trips (mean ± std)	Rush-hour trips (mean ± std)
Time (min) (free flow time)	7.21 ± 5.92	6.62 ± 5.94	7.03 ± 6.38
Time (min) (Actual travel time)	10.61 ± 9.17	12.31 ± 12.85	13.39 ± 13.31
Distance (m)	5958 ± 5937	5419 ± 5952	6049 ± 6895
Number of intersections	3.68 ± 3.84	3.41 ± 3.82	3.62 ± 4.08
Route directness index	1.56 ± 0.88	1.59 ± 1.01	1.56 ± 0.85
Turns statistics			
Left turns	1.27 ± 2.01	1.17 ± 1.90	1.22 ± 1.96
Right turns	1.22 ± 1.86	1.11 ± 1.84	1.13 ± 1.83
Total normal turns	2.49 ± 2.96	2.28 ± 2.91	2.35 ± 2.95
Sharp left turns	1.77 ± 2.03	1.76 ± 2.13	1.80 ± 2.14
Sharp right turns	1.92 ± 2.12	1.83 ± 2.16	1.88 ± 2.17
Total sharp turns	3.69 ± 3.40	3.59 ± 3.60	3.68 ± 3.58
Total all turns	6.18 ± 5.54	5.87 ± 5.77	6.03 ± 5.75
% of route based on road type			
% distance on primary roads	0.38 ± 0.35	0.32 ± 0.35	0.33 ± 0.35
% distance on secondary roads	0.27 ± 0.30	0.27 ± 0.30	0.25 ± 0.29
% distance on tertiary roads	0.35 ± 0.33	0.41 ± 0.35	0.42 ± 0.35
Overlap with corresponding optimal paths			
% of shared length with shortest paths	0.50 ± 0.38	0.51 ± 0.38	0.52 ± 0.38
% of shared length with fastest paths	0.52 ± 0.39	0.52 ± 0.39	0.54 ± 0.39
Centrality			
Betweenness centrality	0.00333 ± 0.00640	0.00319 ± 0.00635	0.00354 ± 0.00697
Closeness centrality	0.00830 ± 0.00044	0.00828 ± 0.00048	0.00829 ± 0.00048
Information centrality	1.0589-06 ± 8.2813-08	1.0529-06 ± 8.4597-08	1.0500-06 ± 8.4766-08
Degree centrality	2.9594-05 ± 2.5348-06	2.9458-05 ± 2.6158-06	2.9396-05 ± 2.5988-06

Length

The trips during the rush-hours are slightly longer than night trips, while trips during daytime are the shortest ones. The fact that the night trips and rush-hour trips are very close together in terms of average length seems a bit odd at first, considering that these two groups are more distinct from each other than for example the daytime and rush-hour group. Since there is less traffic during the night one could expect trips at these times to be shorter, since taking more direct routes than during daytime is not much constrained by traffic conditions. However, the fastest routes are usually not the most direct ones and taxi drivers seem to minimize travel time and not travel distance.

Actual travel time

The mean values for the actual travel time show a pattern which could be expected. The trips took the least amount of time (10.61 minutes) during the night and the most (13.39) during the rush-hour period. The value for daytime trips lies in between the other two groups (12.31). Thus, the amount of traffic seems to directly influence the travel time. However, since the trip purposes are unknown, the influence of different natures of the trips on travel times is unclear. But since the differences in travel distance are much smaller than the differences in actual travel time, it is reasonable to attribute most differences in actual travel times to varying traffic conditions.

Free-flow travel time

Looking at the numbers for the distance variable, it gets clear that the difference in the length of the trips can only partly explain the differences in free-flow travel time. On one hand, the night trips are on average more than 500 meters longer than the day trips. On the other hand, the rush-hour trips are even longer while having lower free-flow travel time values than the nighttime trips. So different travel distances only account for a small part of the differences in the free-flow travel times. The remaining difference can be explained by looking at the actual travel times. These show a more expected pattern (see above).

Number of intersections

The number of intersections along the chosen path does not vary a lot between the groups. The differences are not relevant, in the case of the daytime and rush-hour trips, the difference is not even statistically significant. Figure 19 displays the results with boxplots.

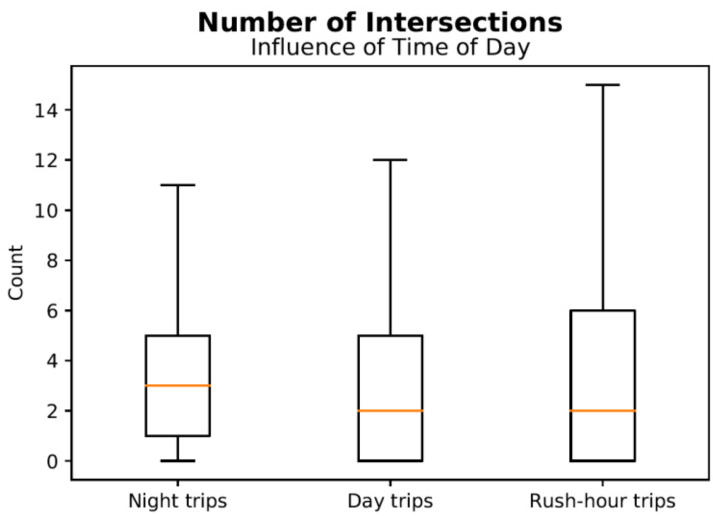

Figure 19: Number of intersections split by time of the day.

Route directness index

The RDI values are almost the same for all groups (0.63 for daytime trips, 0.64 for rush-hour and nighttime trips). This contradicts the intuition that trips are more direct during times with less traffic than during rush-hour where there is a lot of traffic, which leads to people looking for less direct routes to circumnavigate traffic jams for example (minimizing traffic time at the cost of increased travel distance). However, no such effects can be seen in the results.

Turns

As is the case for most variables, the variation in the number of turns is relatively small (they are significantly smaller than the variations between computed and observed paths). Again, as for the theoretical free-flow travel time, the results are somewhat unexpected and run counter to the intuition that trips during the night are more direct than during the day. Again, the high number of turns relative to the other groups can be explained to some extent with the differences in travel distance – the longer a trip, the more turns can be expected. However, nighttime trips have more turns than rush-hour trips even though they are generally shorter. Considering that nighttime trips have the lowest actual travel times while being longer than daytime trips and of almost similar length as the rush-hour trips, one can assume that optimal paths usually involve minimizing travel times at the expense of minimizing travel distance. Figure 20 shows the results as boxplots.

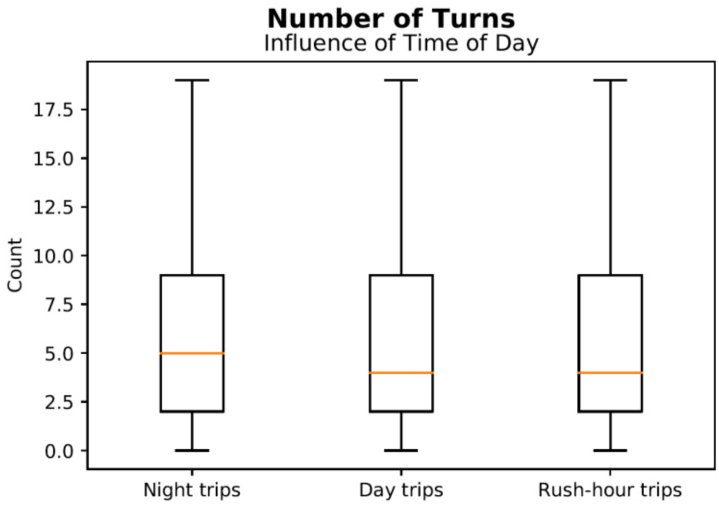

Figure 20: Number of turns per time of the day.

Road type

Predominant in the rush-hour group is the tertiary road category, while the primary road category is predominant in the nighttime group. The results for this variable indicate that primary roads are preferable under conditions with less traffic. During daytime it might be faster to switch to lower order roads (especially tertiary roads), where the speed limits are lower but at the same time there is probably less traffic than on primary roads. The results, visualized with boxplots, can be seen in Figure 21, Figure 22 and Figure 23.

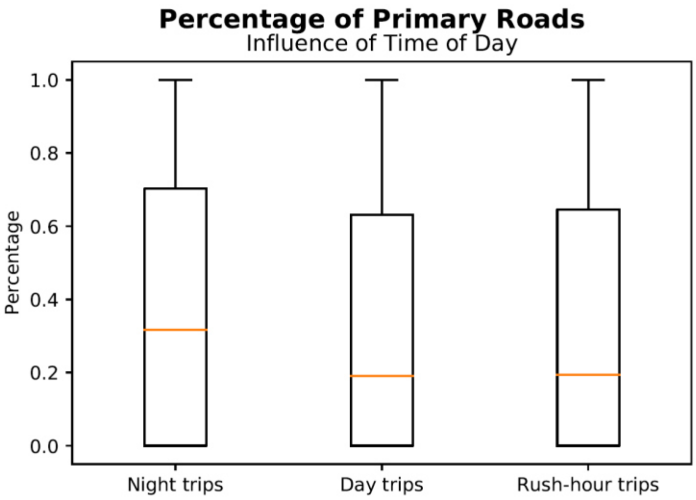

Figure 21: Use of primary roads per time of the day.

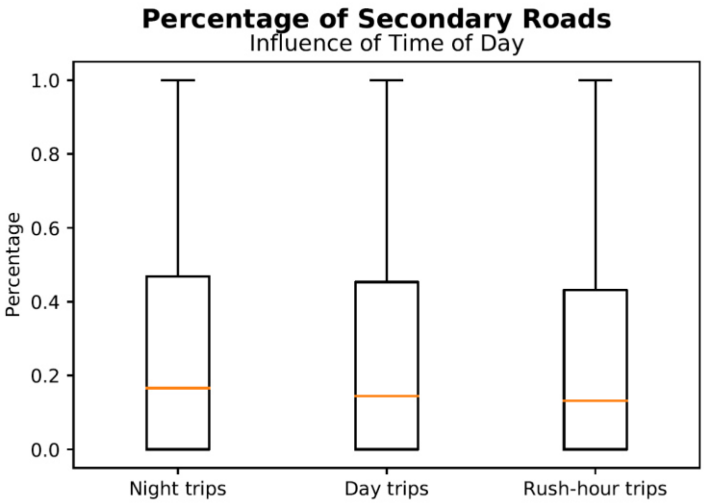

Figure 22: Use of secondary roads per time of the day.

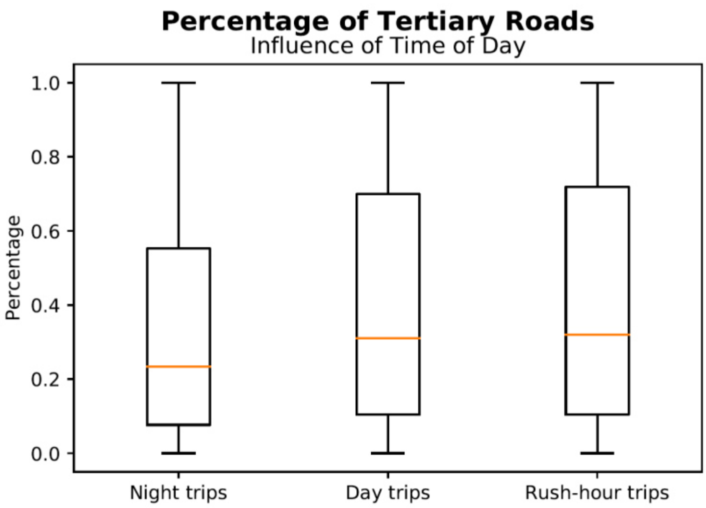

Figure 23: Use of tertiary roads per time of the day.

Percentage of shared length

The differences in PSL while being statistically significant most of the time, are rather small. The overlap with fastest paths is bigger than the overlap with shortest paths for all three groups of observed paths. Surprisingly, PSL values get higher from nighttime to daytime as well as from daytime to rush-hour.

Centrality indices

The biggest differences can be found for betweenness centrality between day trips and rush-hour trips, and for information centrality between rush-hour and night trips. The chosen routes seem to be central during rush-hour, while not being as critical in terms of information centrality compared to nighttime trips.

General observations and summary

While still being statistically significant, t-statistics for most variables are smaller than is the case for the differences between observed and optimal time and distance paths. Thus, the influence of time of day on route choice behavior does not seem to be very strong (however, there are significant differences in space as we will see in the case study-section). The most articulate differences can be found for the percentage of the chosen road types and for the actual travel times. During nighttime taxi drivers prefer primary roads and they arrive faster at the destination. During the rush-hour period on the other hand, tertiary roads are the most popular streets among drivers and the trips take more time to complete. This is of course due to the varying amount of traffic. These results suggest that, while the trips during hours with a high amount of traffic take more time than trips of similar length during nighttime, they would be even slower when driving more on primary and secondary roads at the expense of driving on tertiary roads. Or in the other way around, the difference in the amount of traffic between night and daytime is relatively higher for primary roads than for tertiary roads. This makes primary roads less appealing to taxi drivers during times with a lot of traffic and thus, they switch to lower order roads to circumvent the traffic.

Table 14: t-test statistics and corresponding p-values for the differences between observed paths grouped by time of day. Bold values indicate differences significant at the 0.05 significance level (n = 20'000).

Variables	Night vs. day trips	Day vs. rush-hour trips	Rush-hour vs. night trips
Time (free flow)	**10.00 (0.000)**	**-6.65 (0.000)**	**-2.98 (0.003)**
Time (actual travel time)	**-15.18 (0.000)**	**-8.31 (0.000)**	**24.34 (0.000)**
Distance	**9.05 (0.000)**	**-9.78 (0.000)**	1.42 (0.154)
Number of intersections	**7.12 (0.000)**	**-5.37 (0.000)**	-1.53 (0.127)
Route directness index	**3.56 (0.000)**	**2.10 (0.036)**	**-4.48 (0.000)**
Turns statistics			
Left turns	**5.27 (0.000)**	**2.61 (0.009)**	**-2.66 (0.008)**
Right turns	**5.58 (0.000)**	**4.74 (0.000)**	-0.87 (0.386)
Total normal turns	**7.04 (0.000)**	**4.72 (0.000)**	**-2.29 (0.022)**
Sharp left turns	0.47 (0.636)	1.40 (0.161)	-1.83 (0.067)
Sharp right turns	**4.05 (0.000)**	1.88 (0.060)	**-2.14 (0.032)**
Total sharp turns	**2.76 (0.006)**	0.32 (0.750)	**-2.38 (0.017)**
Total turns	**5.35 (0.000)**	**2.67 (0.008)**	**-2.65 (0.008)**
% of route based on road type			
% distance on primary roads	**16.52 (0.000)**	**-3.50 (0.000)**	**-12.87 (0.000)**
% distance on secondary roads	**2.21 (0.027)**	**4.49 (0.000)**	**-6.73 (0.000)**
% distance on tertiary roads	**-18.57 (0.000)**	-0.32 (0.748)	**18.85 (0.000)**
Overlap with corresponding optimal paths			
% of shared length with shortest paths	**-3.09 (0.002)**	-1.29 (0.198)	**9.09 (0.000)**
% of shared length with fastest paths	0.95 (0.344)	**-4.29 (0.000)**	**3.35 (0.001)**
Centrality			
Betweenness centrality	**2.23 (0.025)**	**-5.16 (0.000)**	**3.01 (0.003)**
Closeness centrality	**3.34 (0.000)**	-0.12 (0.902)	**-3.31 (0.000)**
Information centrality	**7.18 (0.000)**	**3.34 (0.000)**	**-10.55 (0.000)**
Degree centrality	**5.28 (0.000)**	**2.38 (0.017)**	**-7.72 (0.000)**

5.2.2 Day of the week

Next, the trips have been split into two groups depending on the day of the week: Trips conducted on Saturdays or Sundays build the weekend group, all other trips compose the weekday group. The weekday group consists of 1'806'598 trips and the weekend group consists of 785'837 paths.

The purpose of the comparison between these two groups is to find out what time related factors influence the route characteristics more. Is time of the day or the day of the week? Unsurprisingly, the differences between weekday and weekend are quite similar to the differences between day-time/rush-hour and nighttime trips. Therefore, the analysis is limited here to the most important parts and the differences to the comparisons in the previous groups are highlighted.

Table 15: Route characteristics of weekday and weekend trips.

Variables	Weekday trips (mean ± std)	Weekend trips (mean ± std)
Time (min) (free flow time)	6.75 ± 6.02	6.91 ± 5.76
Time (min) (Actual travel time)	12.19 ± 12.13	10.90 ± 11.67
Distance (m)	5494 ± 6045	5680 ± 5855
Number of intersections	3.46 ± 3.83	3.48 ± 3.76
Route directness index	1.59 ± 1.05	0.63 ± 0.55
Turns statistics		
Left turns	1.18 ± 1.96	1.23 ± 1.96
Right turns	1.07 ± 1.78	1.15 ± 1.84
Total normal turns	2.25 ± 2.92	2.38 ± 2.92
Sharp left turns	1.72 ± 2.08	1.75 ± 2.05
Sharp right turns	1.83 ± 2.16	1.88 ± 2.13
Total sharp turns	3.55 ± 3.53	3.63 ± 3.45
Total all turns	5.80 ± 5.69	6.01 ± 5.57
% of route based on road type		
% distance on primary roads	0.32 ± 0.35	0.35 ± 0.35
% distance on secondary roads	0.27 ± 0.29	0.27 ± 0.30
% distance on tertiary roads	0.41 ± 0.35	0.38 ± 0.34
Overlap with corresponding optimal paths		
% of shared length with shortest paths	0.52 ± 0.38	0.51 ± 0.38
% of shared length with fastest paths	0.53 ± 0.39	0.53 ± 0.39
Centrality		
Betweenness centrality	0.00322 ± 0.00643	0.00325 ± 0.00618
Closeness centrality	0.00829 ± 0.00047	0.00828 ± 0.00046
Information centrality	1.0539-06 ± 8.3775-08	1.0543-06 ± 8.4513-08
Degree centrality	2.949-05 ± 2.611-06	2.9566-05 ± 2.6122-06

Length

As it has been the case with the results from the previous factor, trips during time where there was probably less traffic, namely on weekends, are longer (5.68 km) than the trips during weekdays (5.49 km).

Actual travel time

Weekday trips take significantly more time than weekend trips, while being shorter. These differences can therefore be explained again with the varying amount of traffic.

Free-flow travel time

The weekend trips would theoretically take longer under free-flow conditions than the weekday trips. For actual travel times it is exactly the other way around as we have just seen before. Thus, theoretically optimal paths are usually not the best routes when adding traffic to the equation.

Number of intersections

The average number of intersections along the path is very similar for the two groups (3.46 vs. 3.48). The difference is not relevant let alone statistically significant. Since weekend trips are longer however, this means that travelers drive along roads with relatively fewer intersections during the week and with relatively more intersections during the weekend.

Route directness index

The RDI value for both groups equals 0.63. Thus, there is no visible influence of the day of week on the directness of the chosen routes.

Turns

There are more turns on weekends than on weekdays. However, the trips are also longer on weekends, which is most likely the reason for the slight differences between the groups.

Road type

Again, we can see that more traffic seems to positively correlate with the usage of tertiary roads, and conversely, less traffic is positively correlated with the use of primary streets.

Percentage of shared length

There are no apparent differences for the PSL-variable. The trips overlap a little less with their corresponding shortest paths on weekends than on weekdays. This could mean that under conditions without a lot of traffic, less direct routes are usually faster in terms of travel time. However, as indicated before, the differences are marginal.

Centralities

The differences of the centrality indices between the two groups of trips are not relevant and except for the degree centrality index, they are also not statistically significant.

General observations and summary

The main difference between the weekday and weekday groups is most likely again the amount of traffic, which is smaller for weekends than during the week. Therefore, it is not surprising that the differences between weekdays and weekends show similar patterns than the differences between daytime/rush-hour trips and nighttime trips. However, the patterns are less pronounced, which makes sense as the variation of traffic during the day superposes the variation between weekends and weekdays. Additionally, the latter is most likely also bigger than the former. Hence, standard deviations are generally bigger here than in the time of the day-groups from before.

Table 16: t-test statistics and corresponding p-values for the differences between observed paths grouped by day of week. Bold values indicate differences significant at the 0.05 significance level (n = 20’000).

Variables	Weekday vs. weekend trips
Time (free flow)	**-2.66 (0.008)**
Time (actual travel time)	**10.7 (0.000)**
Distance	**-3.13 (0.001)**
Number of intersections	-0.59 (0.558)
Route directness index	**-2.02 (0.044)**
Turns statistics	
Left turns	**-2.41 (0.016)**
Right turns	**-4.21 (0.000)**
Total normal turns	**-4.23 (0.000)**
Sharp left turns	-1.76 (0.078)
Sharp right turns	**-2.43 (0.015)**
Total sharp turns	**-2.53 (0.011)**
Total turns	**-3.76 (0.000)**
% of route based on road type	
% distance on primary roads	**-7.53 (0.000)**
% distance on secondary roads	**-3.74 (0.000)**
% distance on tertiary roads	**10.6 (0.000)**
Overlap with corresponding optimal paths	
% of shared length with shortest paths	**3.09 (0.002)**
% of shared length with fastest paths	1.71 (0.088)
Centrality	
Betweenness centrality	-0.54 (0.591)
Closeness centrality	1.75 (0.079)
Information centrality	-0.45 (0.649)
Degree centrality	**-2.91 (0.004)**

5.2.3 Weather

Lastly, the trips have been split according to weather conditions present when the trip has been conducted. There are two groups, trips under ‘good’ weather conditions and trips under ‘bad’ weather conditions. The former group consists of the conditions ‘Clouds’ and ‘Clear’ while the latter consists of the conditions ‘Rain’ and ‘Snow’. The bad weather group consists of 380’903 and there are 1’734’218 trips in the good weather group. When the weather conditions change during a trip, it is put into both groups. Since the sampling rate of the weather dataset is one hour and a trip took about 12 minutes on average this should not affect the results too much.

The results show that the influence of weather conditions on the route characteristics is minimal. Only three variables are statistically significantly different: The actual travel time, the travel distance and the number of intersections (Table 17 and Table 18). This is due to varying trip lengths in the samples. The trips in the bad weather group are a little bit longer on average and therefore they also take a little more time and have a bit more intersections along their way. Therefore, the weather conditions do not have any noteworthy influence on the characteristics of the chosen routes.

Table 17: Route characteristics of good and bad weather trips.

Variables	Good weather trips (mean ± std)	Bad weather trips (mean ± std)
Time (min) (free flow time)	6.99 ± 6.12	7.10 ± 6.25
Time (min) (Actual travel time)	12.28 ± 11.7	12.75 ± 14.20
Distance (m)	5818 ± 6327	6032 ± 6559
Number of intersections	3.53 ± 3.96	3.62 ± 3.97
Route directness index	1.61 ± 0.91	1.59 ± 2.92
Turns statistics		
Left turns	1.24 ± 1.98	1.24 ± 2.01
Right turns	1.15 ± 1.86	1.16 ± 1.85
Total normal turns	2.39 ± 2.96	2.40 ± 2.97
Sharp left turns	1.77 ± 2.03	1.79 ± 2.08
Sharp right turns	1.88 ± 2.14	1.90 ± 2.14
Total sharp turns	3.65 ± 3.43	3.69 ± 3.49
Total all turns	6.04 ± 5.61	6.09 ± 5.67
% of route based on road type		
% distance on primary roads	0.35 ± 0.35	0.35 ± 0.35
% distance on secondary roads	0.26 ± 0.30	0.26 ± 0.30
% distance on tertiary roads	0.39 ± 0.34	0.39 ± 0.34
Overlap with corresponding optimal paths		
% of shared length with shortest paths	0.51 ± 0.38	0.51 ± 0.38
% of shared length with fastest paths	0.52 ± 0.39	0.52 ± 0.39
Centrality		
Betweenness centrality	0.00339 ± 0.00633	0.00333 ± 0.00606
Closeness centrality	0.00829 ± 0.00047	0.00829 ± 0.00047
Information centrality	1.0522-06 ± 8.4436-08	1.0525-06 ± 8.4244-08
Degree centrality	2.9493-05 ± 2.6080-06	2.9449-05 ± 2.5777-06

Bad weather (rain and snow) has generally a negative impact on the traffic speed (Edwards 1999; Oh, Shim, and Cho 2002; Akin, Sisiopiku, and Skabardonis 2011). Therefore, one could expect the bad weather trips to take more time than good weather trips (assuming similar trip lengths). However, this is not represented in the data. One reason for this could be that lower speeds get canceled out by less traffic during adverse weather conditions. In fact, several studies report a negative correlation between rain/snow and the amount of traffic (Keay and Simmonds 2005; Datla and Sharma 2010). Therefore, taxi drivers may drive a little bit slower during harsh weather, but the trips do not take longer because there are fewer delays due to traffic.

Table 18: t-test statistics and corresponding p-values for the differences between observed paths grouped by weather conditions. Bold values indicate differences significant at the 0.05 significance level (n = 20'000).

Variables	Good weather vs. bad weather trips
Time (free flow)	-1.70 (0.090)
Time (actual travel time)	**-3.59 (0.000)**
Distance	**-3.32 (0.001)**
Number of intersections	**-2.28 (0.023)**
Route directness index	-0.61 (0.541)
Turns statistics	
Left turns	-0.51 (0.607)
Right turns	-0.61 (0.542)
Total normal turns	-0.73 (0.467)
Sharp left turns	-1.02 (0.308)
Sharp right turns	-0.56 (0.577)
Total sharp turns	-0.95 (0.342)
Total turns	-0.97 (0.334)
% of route based on road type	
% distance on primary roads	-0.30 (0.762)
% distance on secondary roads	-0.09 (0.365)
% distance on tertiary roads	0.92 (0.359)
Overlap with corresponding optimal paths	
% of shared length with shortest paths	0.99 (0.324)
% of shared length with fastest paths	-0.04 (0.965)
Centrality	
Betweenness centrality	0.95 (0.342)
Closeness centrality	-0.3 (0.760)
Information centrality	-0.31 (0.759)
Degree centrality	1.67 (0.094)

5.3 Case studies

In the previous chapter, trips have been characterized using eight variables. Then, observed trips have been compared with shortest and fastest trips. Last, the influence of three factors on these route characteristics has been investigated. There is another variable, which has not really been incorporated into the analysis so far: Geography. Different areas in the city of Vienna have distinct characters. The way the city is built differs from one part of the city to another. The inner city for sure has distinctive characteristics with its historic buildings which are hundreds of years old compared to newer parts of the city such as the area around the Austria Center Vienna with its modern high-rise buildings. More importantly, the street network configuration differs considerably in space: There are a lot of narrow, one-way streets in the historic center of the city. On the other hand, other districts such as Florisdorf and Donaustadt have broader streets and less one-way and turn restrictions and can thus be characterized as more car friendly. Therefore, it is likely that trips in different areas of the city also exhibit different route characteristics. The question of course is, whether these are only due to differences in the street network configuration. For example, if an area has a lot of long and straight streets, which intersect each other only in 90-degree angles then one could expect less turns per kilometer compared to trips in an area with a high street density and circuitous routes. However, there is also the possibility that street network characteristics directly influence route choice behavior. Maybe it is easier to find the fastest or most direct route in areas with a smaller density of streets than in areas with a higher density and a lot of restrictions.

The incorporation of geography into the analysis could reveal differences in route choice behavior, which are not represented in the results of the last chapters. Two trips can have similar route characteristics but that does not necessarily mean that they overlap in space. In fact, they might not share one single road segment but for some reason they have the same number of turns and intersections, are of similar length and the duration of the trip is also equal. Looking at the trips on a map and compare them could reveal such differences which would otherwise remain hidden, when looking only at the results from the last two chapters. This is especially interesting for the analysis of the influence of external factors. As we have seen, there are not always pronounced differences in route characteristics between separate groups of paths. Looking at the results from the previous chapter, we can see that this is true especially for the bad and good weather groups. However, there is still the possibility that taxi drivers are forced to choose different routes when it is snowing compared to clear blue sky-weather conditions.

Directional effects are another aspect, which can only be investigated by looking at the trips in space. As already mentioned in chapter 4, it will be investigated if directional effects similar to those Manley, Addison, and Cheng (2015) have reported, are visible in the present dataset as well. Next, the influence of street network centrality is also investigated best by looking at the trips mapped in space. Only then it is possible to see whether taxi drivers prefer routes with high betweenness centrality values over routes with low betweenness centrality values for example. Lastly, it will be analyzed if taxi drivers tend to avoid certain areas of the city and drive around rather than through them.

Three groups of case studies should help to analyze all of this. In the first group, the influence of external factors on differences in route choice behavior is investigated by taking the same number of trips going from an origin area A to a destination area B from separate groups of observed trips, e.g. 100 trips from A to B carried out under clear weather conditions and 100 trips also from A to B carried out under harsh weather conditions. As described in chapter 4, the differences are calculated on a road segment basis first (e.g. number of times a road segment has been used in the bad weather group minus the number of times a road segment has been used in the good weather group). Then LISA is applied to identify areas which are preferred by drivers on only one weather condition. The same method is applied to analyze the influence of time of the day and day of the week.

In the second group, directional effects are analyzed using the same approach just described above. Only this time, one group consists of trips going from A to B and the other group consists of trips going from B to A.

In the third and last group, there are eight individual case studies which help to analyze the influence of centrality. The first two show whether taxi drivers prefer roads with high betweenness centrality. The next two show if they avoid areas with low information centrality or if it is the other way round and they prefer areas with high information centrality values. Case study six and seven do the same but with closeness centrality. Last, case studies five and eight are used to check if trips in areas with high closeness/information centrality have different route characteristics than trips in areas with low closeness/information centrality.

5.3.1 Driver heterogeneity/homogeneity

Before starting with the case studies outlined before, it is important to check how similar the taxi drivers' routes are using three scenarios. The more overlap, the more similar the drivers are.

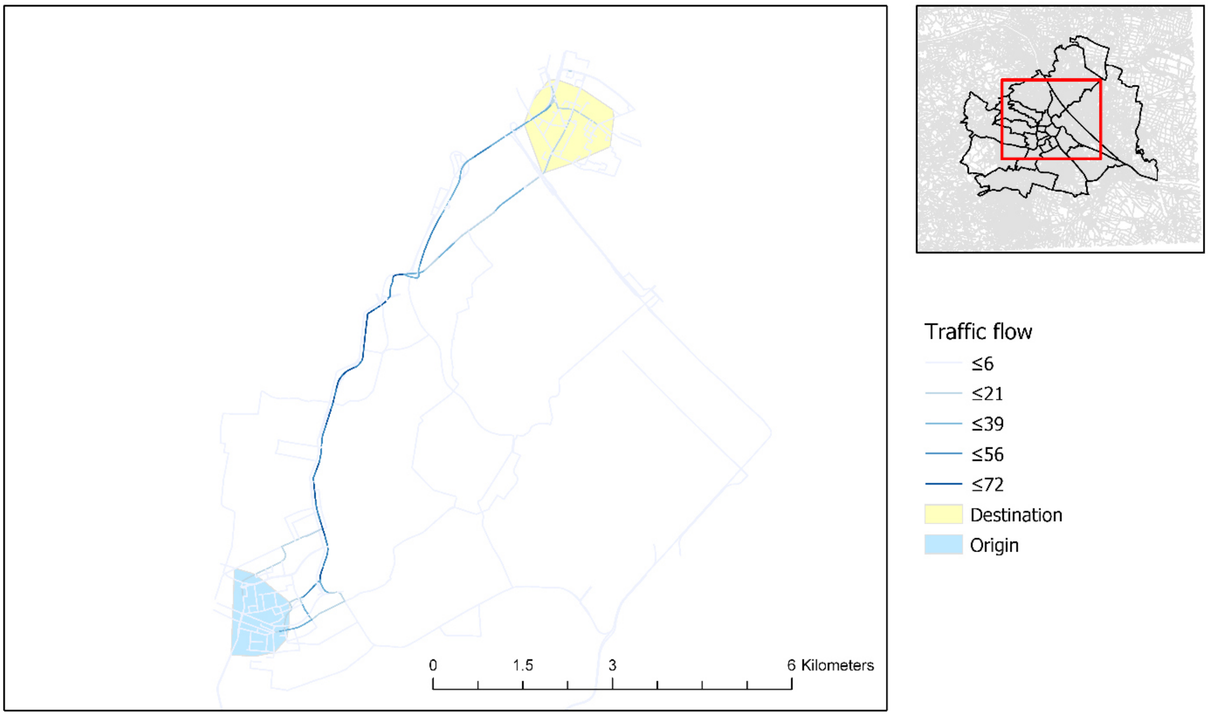

Figure 24: Overlap between observed trips from south to north.

Figure 24 shows trips going through the whole city of Vienna from south to north. The traffic flows show that most drivers reach the destination using the same path, namely highway 227.

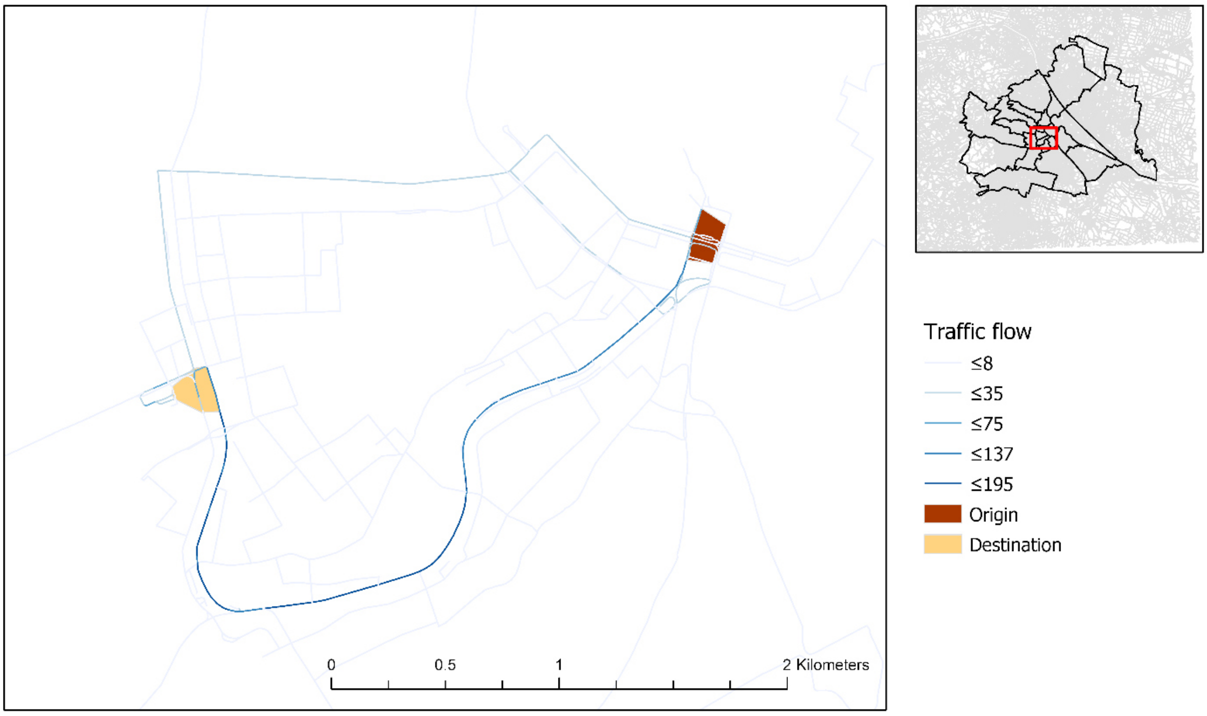

Figure 25: Overlap between observed trips going through the city center.

When going from a location east of the city center to another one west of the city center, taxi drivers mainly choose one out of two paths (see Figure 25). The majority drives around the center using a southern route, a smaller number circumvents the center on a northern route.

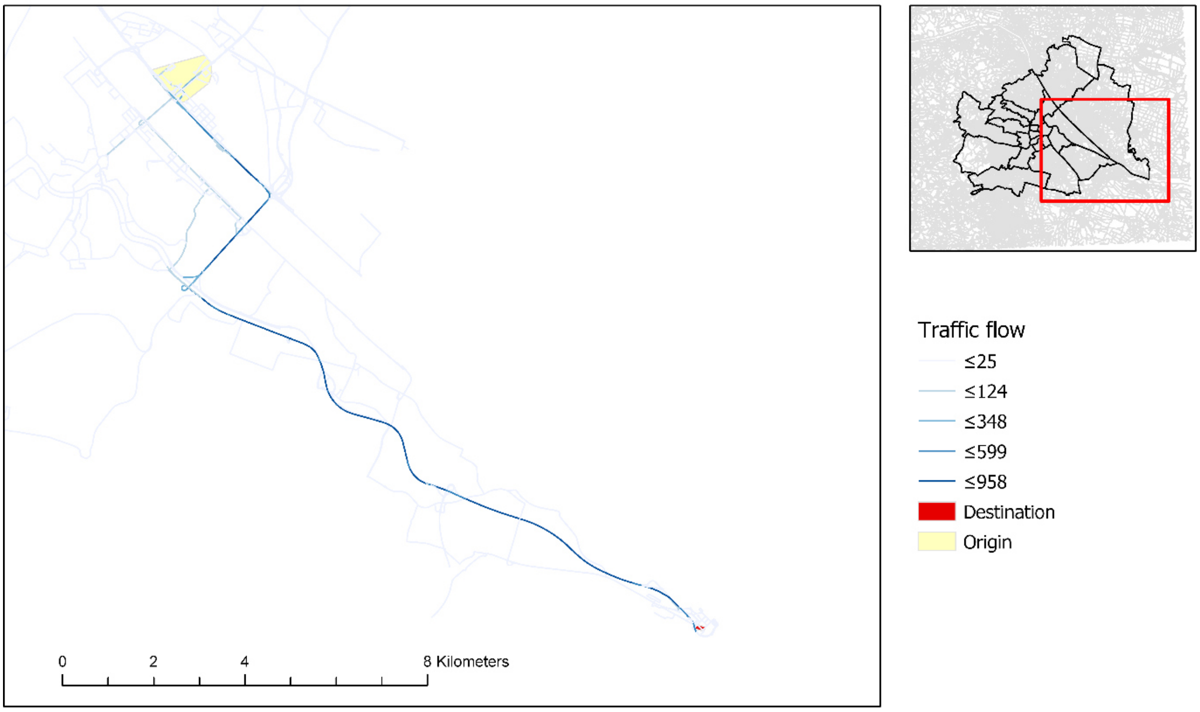

Figure 26: Overlap between trips going from Florisdorf to the airport.

Last, taxi drivers that need to go from the airport on the south to Florisdorf in the north, choose highly overlapping routes (see Figure 26).

In all these three scenarios, the drivers quite homogenous and choose similar routes. Thus, the assumption that they have a similar level of spatial knowledge and experience holds.

5.3.2 Case study group 1: Influence of external factors

Case Study 1 – Am Spitz to Wien Rennweg Bahnhof

In the first case study of group 1, trips going from north to south from Florisdorf over the Danube to a railway station south of the city center are split according to external factors. Figure 27 shows a cluster for rush-hour trips in red going from origin to destination in an anti-clock wise direction mainly using highway 227. During nighttime taxi drivers seem to prefer highway 221 (left hand side in Figure 27), the inner ring around the city center or highway 14 along the Danube approaching the destination in a clock-wise manner. The trips during rush-hour are more direct compared to the trips during nighttime. One explanation could be that during periods with a lot of traffic, it is better to minimize travel distance, as all streets are overloaded, and the average travel speed is low on all alternatives and does not vary much.

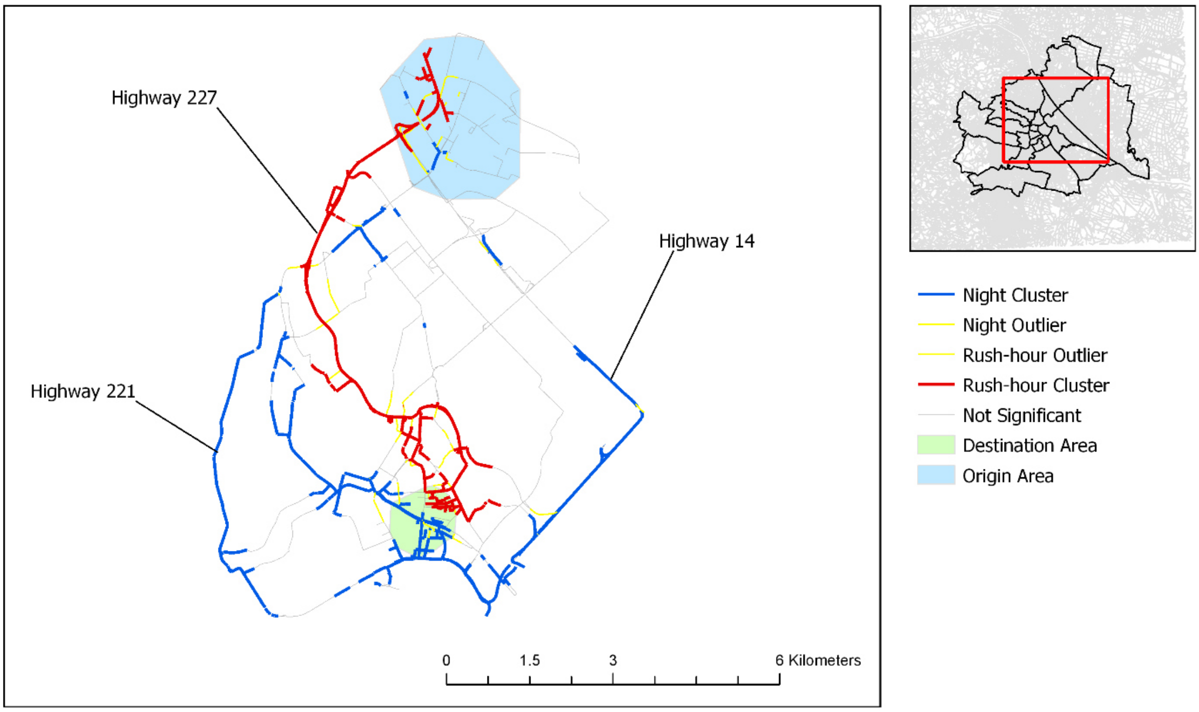

Figure 27: Spatial clustering showing areas with relatively high traffic during the night (blue) and areas with relatively high traffic during rush-hour (n=50 trips per group).

Figure 28 displays the trips clustered according to time of day. The cluster show a similar picture as before: During days with presumably more traffic (weekdays), the chosen routes are quite direct using highway 221 to reach the railway station in the south. However, this time there is no cluster on the right-hand side along the Danube, but there is a cluster of weekend trips which is exactly in between the origin and destination area and thus consists of very direct routes.

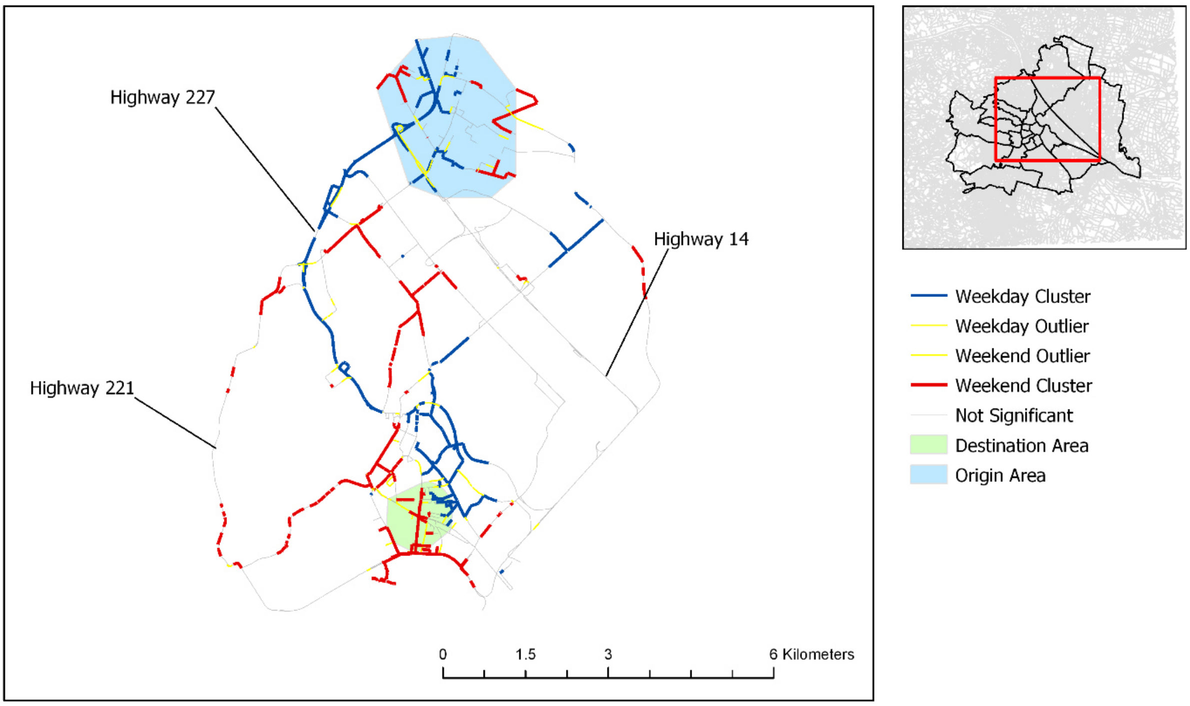

Figure 28: Spatial clustering showing areas with relatively high traffic during weekdays (blue) and areas with relatively high traffic during weekends (n=50).

Looking at the situation in Figure 29, it generally seems that the bad weather trips are clustered more on the left-hand side, while the good weather trips generally are more dominant on the right-hand side and tend to follow along the Danube. Additionally, the picture does not differ very much compared to the other two factors. This is due to a limitation of the methodology applied here. Other variables such as time of the day and day of the week are not held constant. This leads to a distortion of the results. Therefore, differences for the trips regarding these other variables overlay changes that are due to varying weather conditions. In this first case study, this leads to three very similar pictures for all external factors. The conclusion from this is that time of the day has by far the most influence on where chosen routes lead along.

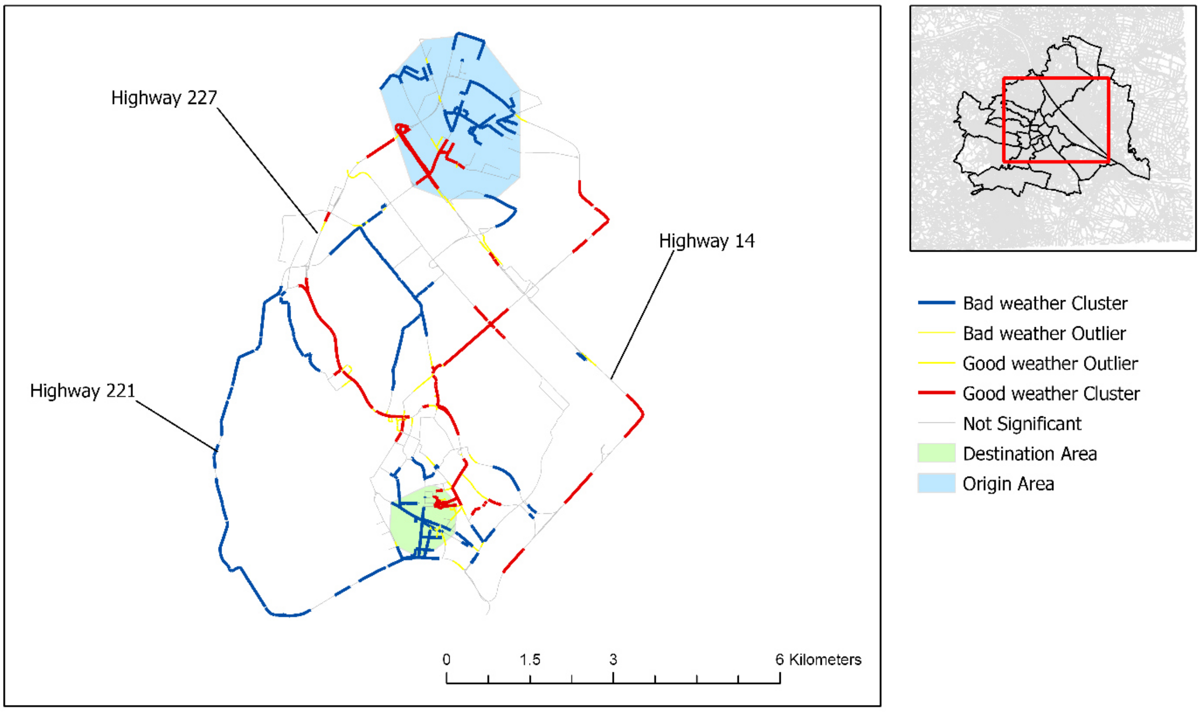

Figure 29: Spatial clustering showing areas with relatively high traffic during bad weather conditions and areas with relatively high traffic during good weather conditions (n=50).

Case Study 2 – Praterstern to Wiener Stadthalle

In the second case study, trips going from the Praterstern in the east going around or through the city center to the Wiener Stadthalle are selected for the analysis. Following the same order as in the previous case study, Figure 30 shows the clustering results when splitting the trips according to time of day. Again, rush-hour trips seem to be a bit more direct than nighttime trips. All trips avoid the city center, but during the nighttime drivers mostly choose the southern route around the center using the inner ring road, while they drive north of the center to the destination during the rush-hour period following along Türkenstrasse – Universitätstrasse – Landesgerichtstrasse. There is also a second nighttime cluster north of the destination area leading along highway 221, which supports the thesis that drivers tend to choose less direct routes during the night, as this minimizes travel time as very high travel speeds can be achieved on highways and other primary roads during free-flow traffic conditions.

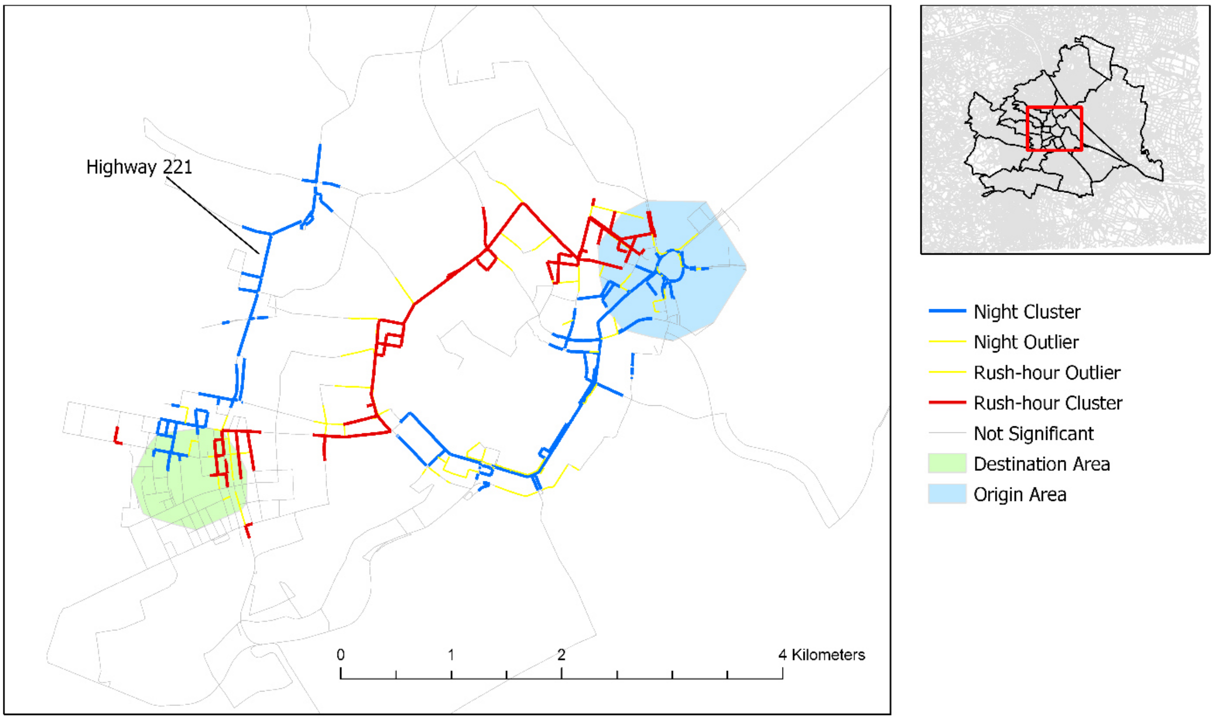

Figure 30: Spatial clustering showing areas with relatively high traffic during the night (blue) and areas with relatively high traffic during rush-hour (n=125).

The weekday and weekend cluster in Figure 31 show a very similar pattern as those in Figure 30. The weekday cluster resembles closely the rush-hour cluster and the weekend cluster is similar to the nighttime cluster. The difference is that in contrast to the nighttime cluster, taxi drivers in the case study usually go on to the highway 'Linke Wienzeile' after they drove around the city center in weekend trips. Also, the cluster around highway 221 is missing.

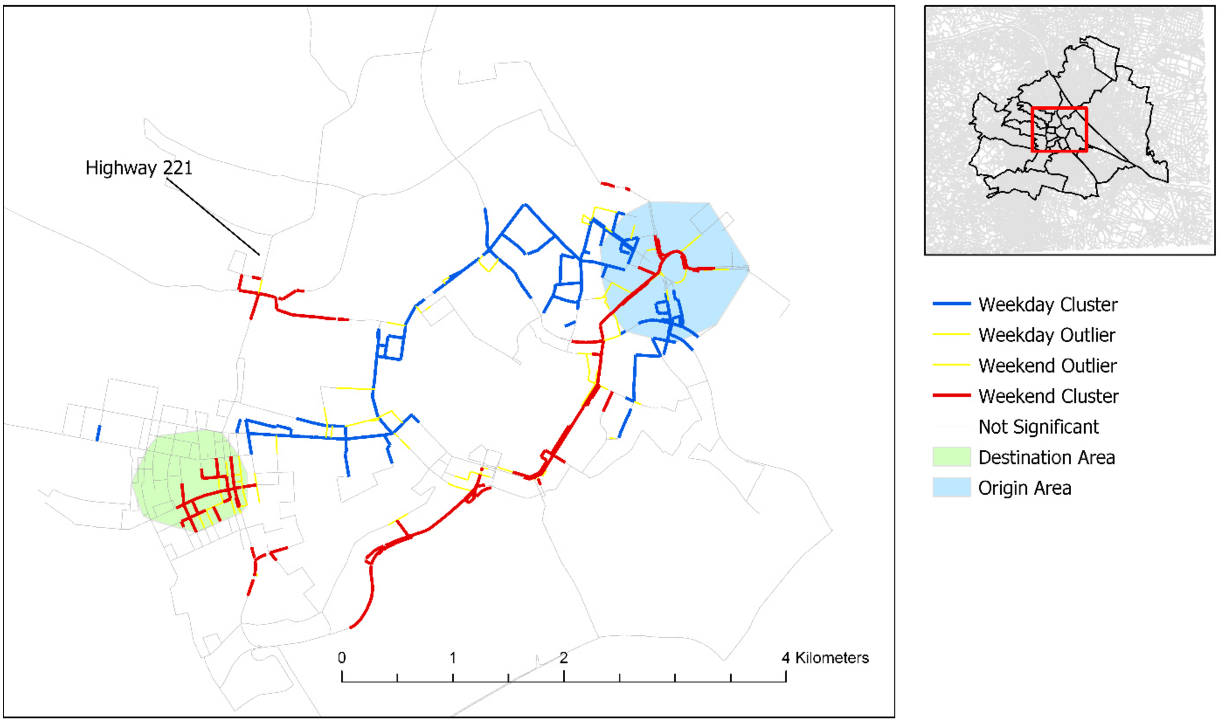

Figure 31: Spatial clustering showing areas with relatively high traffic during weekdays (blue) and areas with relatively high traffic during weekends (n=125).

For the last factor in this case study (Figure 32), there exist again two distinct clusters. Taxi drivers prefer the northern route around the center during good weather conditions and the southern route during bad weather conditions. Again, the most likely explanation for this, is that the effects of varying weather conditions get superimposed by the effects of time of the day.

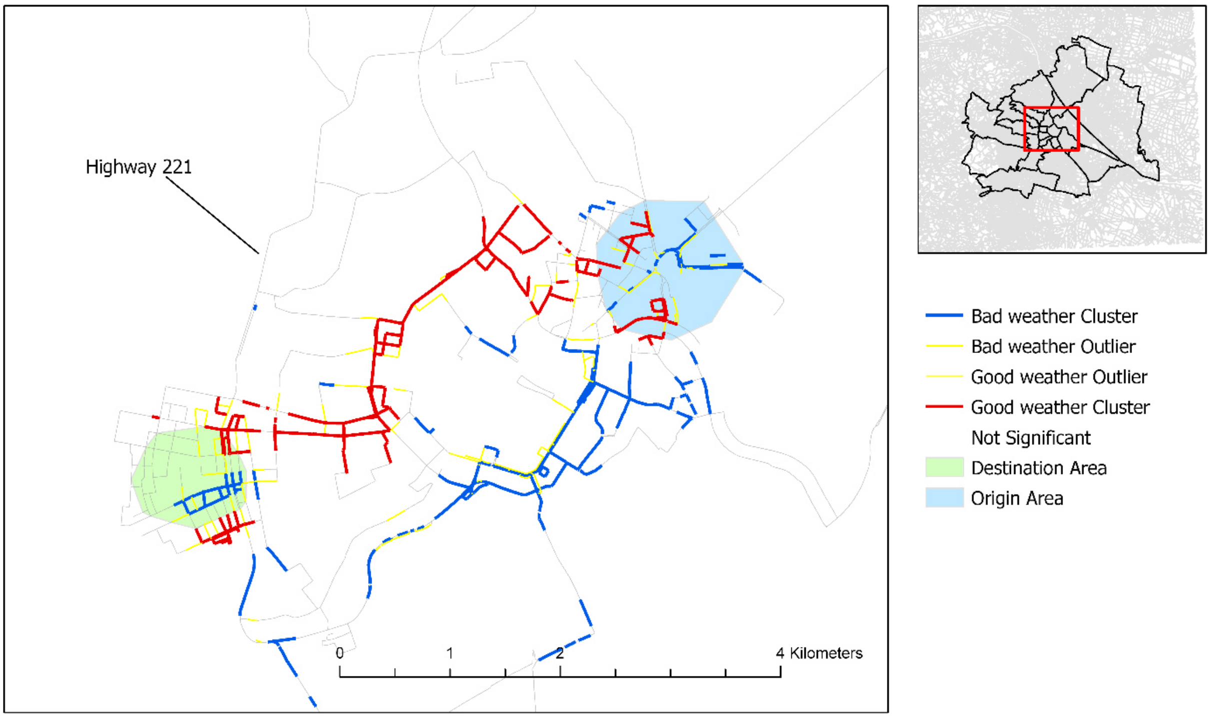

Figure 32: Spatial clustering showing areas with relatively high traffic during bad weather conditions and areas with relatively high traffic during good weather conditions (n=50).

Case Study 3 – Hotel InterContinental to Wiener Stadthalle

The third and last case study where we look at external effects shows less clear patterns than the previous ones. Looking at Figure 33, one can see there are mainly two routes to choose from when trying to get from the area east of the city center to the area on the opposite side of the center, a northern and a more southern route. There is a tendency for taxi drivers to choose the northern route relatively more often during nighttime than during the rush-hour period. Other than that, no cluster can be identified in this case.

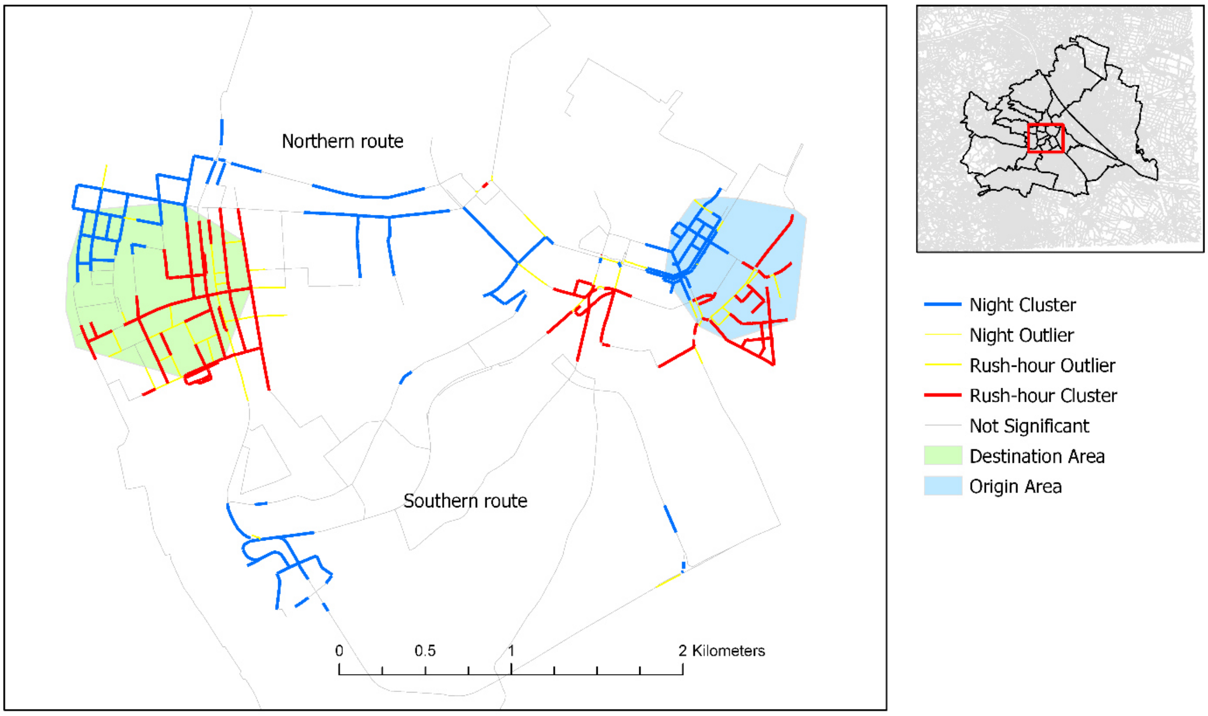

Figure 33: Spatial clustering showing areas with relatively high traffic during the night (blue) and areas with relatively high traffic during rush-hour (n=100).

When comparing weekday and weekend trips shown in Figure 34, it is difficult to identify any meaningful patterns. In contrast to the previous two case studies, there are also hardly any similarities to the nighttime/rush-hour clusters. We can conclude that day of the week has no visible effect on route choice behavior of Viennese taxi drivers in this case study. No street in the middle connecting the two areas has significant values.

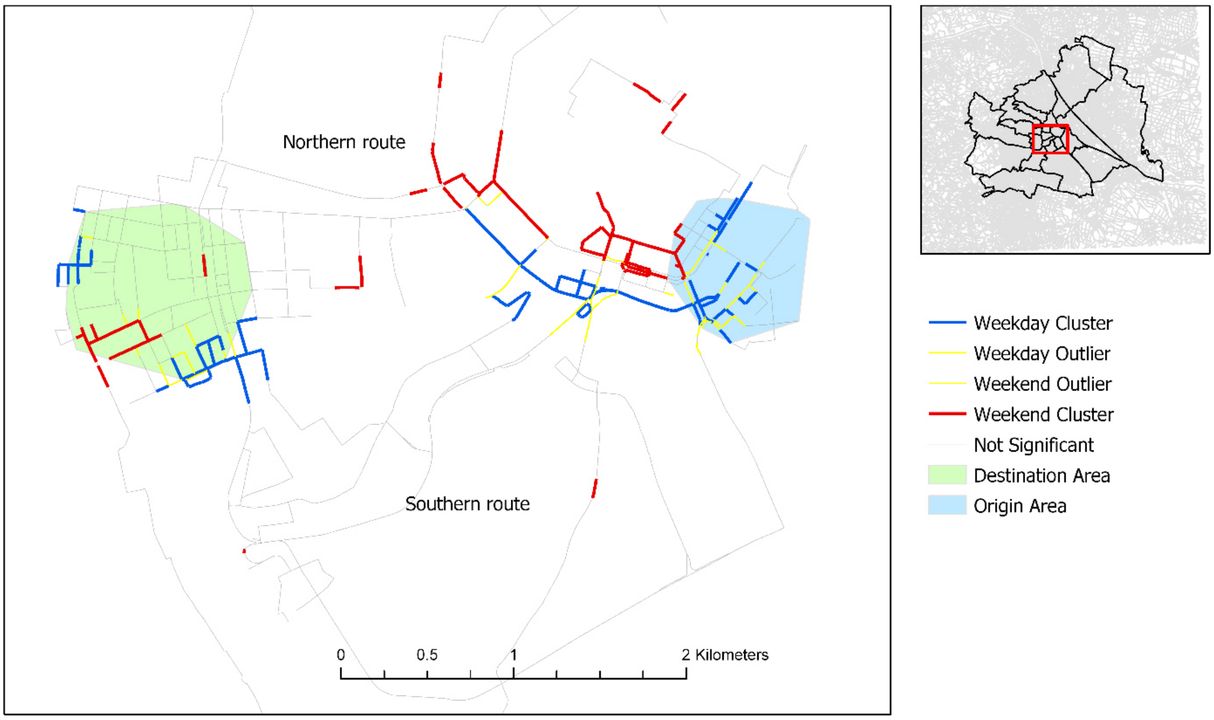

Figure 34: Spatial clustering showing areas with relatively high traffic during weekdays (blue) and areas with relatively high traffic during weekends (n=100).

Last, Figure 35 shows trips clustered according to weather conditions. Interestingly there are way more clusters than in the previous two figures. When the assumption is true that in this scenario the other two variables do not have a significant effect, then these cluster could be a result of differing weather conditions. However, it cannot be ruled out that these are a result of another factor, which has not been identified, or it could be a result of a temporary constraint in the street network. There are significantly fewer trips that have been made during bad weather in the database than trips during good weather. Thus, it could be that exactly on the days where it snowed or rained, a certain road was inaccessible, for example due to a construction site, which would also explain the clusters.

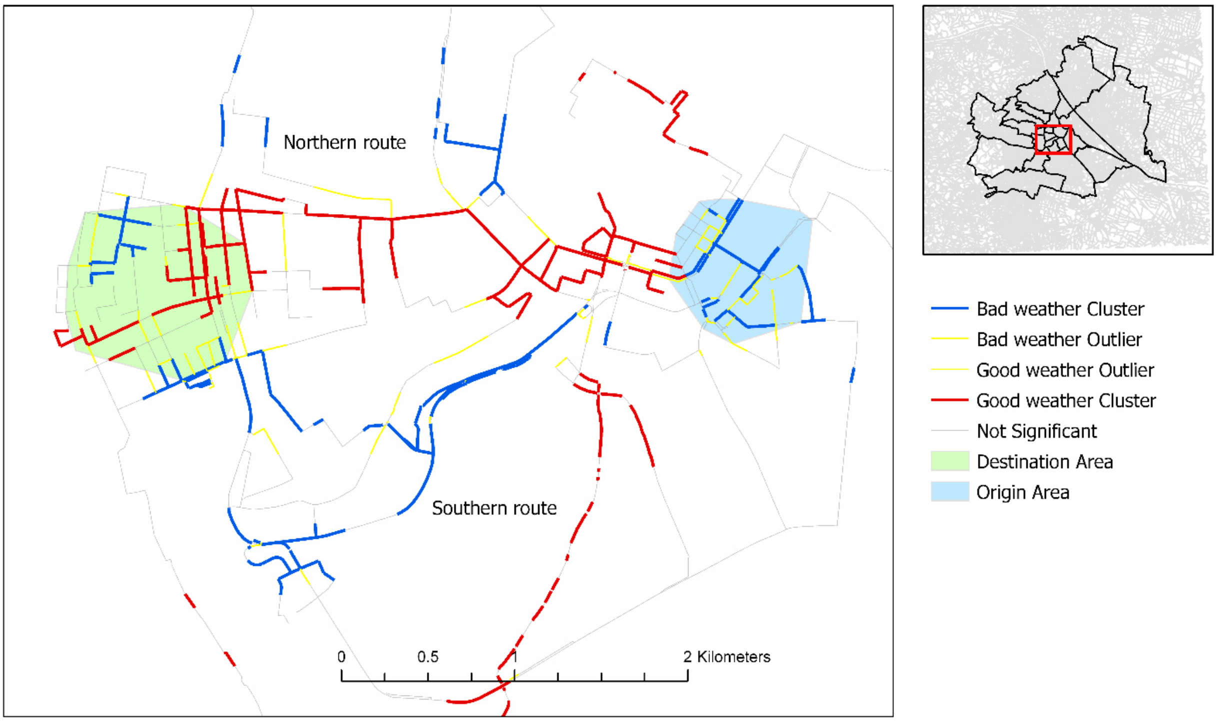

Figure 35: Spatial clustering showing areas with relatively high traffic during bad weather conditions and areas with relatively high traffic during good weather conditions (n=100).

To sum up the results so far, there is evidence for time of the day having a clear influence of the location of observed routes. Furthermore, this influence seems to be much bigger than the effect of the other two factors. The last example showed, that there is a need for more case studies as well as the incorporation of temporary constraints into the analyses to be able to draw clearer conclusions. Additionally, when splitting the trips into groups, the other variables should be held constant. Otherwise there is the possibility the factor under investigation gets superimposed by others, which leads to a distortion of the resulting cluster.

5.3.3 Case study group 2: Directional effects

Case Study 1 – Messe Wien to Hotel Sacher

The origin area of the first case study for directional effects is around Messe Wien, the destination is the area around the five-star hotel Sacher. Figure 36 shows to distinct cluster going from the Praterstern in the north to Karlsplatz in the south. The westbound cluster leads along the inner-city ring road, the eastbound cluster leads along highway 1 and Hintere Zollamtstrasse. However, these directional differences are entirely due to constraints of the street network. The ring-road is a one-way street and thus it is not possible to use it to go from Hotel Sacher to Messe Wien. However, the result indicates that the ring would be preferred over highway 1, as most drivers going from Messe Wien to Hotel Sacher use it instead of the highway.

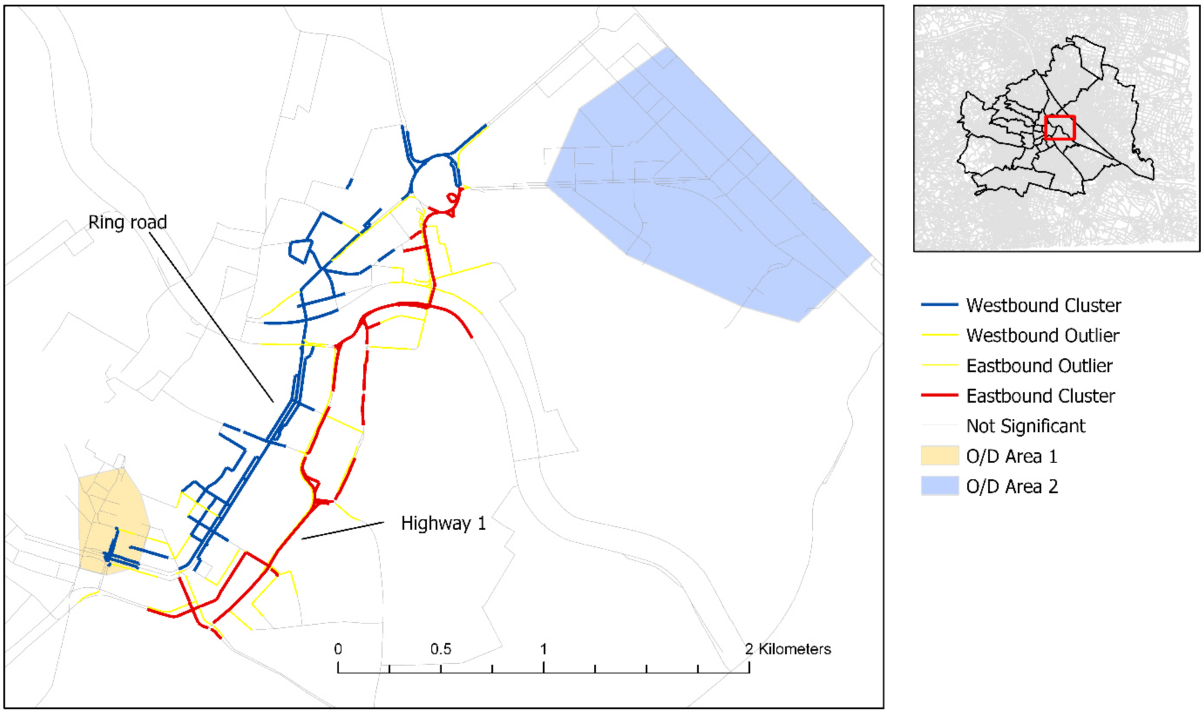

Figure 36: Spatial clustering results from Messe Wien to Hotel Sacher and vice versa indicating areas of high mono-directional traffic flow (n=250).

Case Study 2 – Schloss Schönbrunn to Am Spitz

Looking at the situation in Figure 37, one can see that there are no cluster at all for this second case study. There are some colored road segments along the Gürtel (highway 221), however, no clear pattern is visible. The first option is using highway 221, which directly goes from the origin area to the destination area. Another option is going to the ring road around the city center first and the continue northwards. However, this is only possible when driving from south to north as the ring road is a one-way street. The third option is when starting from the north to go along the Danube first and then to turn right and approach the destination in the south. There are of course many more alternatives. The results show that all three of these routes are used. Nevertheless, most drivers chose highway 221 regardless of travel direction, so that no clear directional effect can be seen in the data.

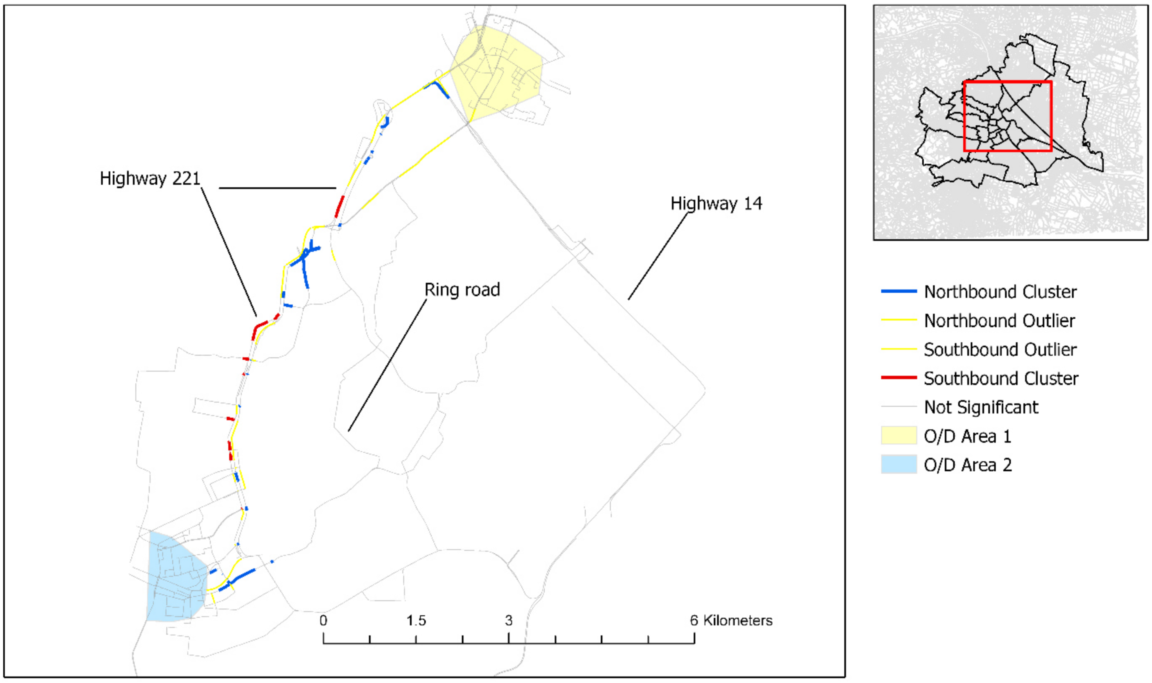

Figure 37: Spatial clustering results from Schloss Schönbrunn to Am Spitz and vice versa indicating areas of high mono-directional traffic flow (n=50).

Case Study 3 – Wien Mitte to Siegmund-Freud-Park

For the third case study on this group, the two OD-regions lie opposite each other along the inner-city ring road. There is a clear eastbound cluster in the north and a westbound cluster in the south. These cluster however, are a consequence of the street network configuration of the city center. As we have seen multiple times now, the ring road around the center is a one-way street and it is only possible to drive clock-wise. Therefore, the results displayed in Figure 38 are not a sign of any directional effects on route choice behavior of taxi drivers but of constraints of the street network.

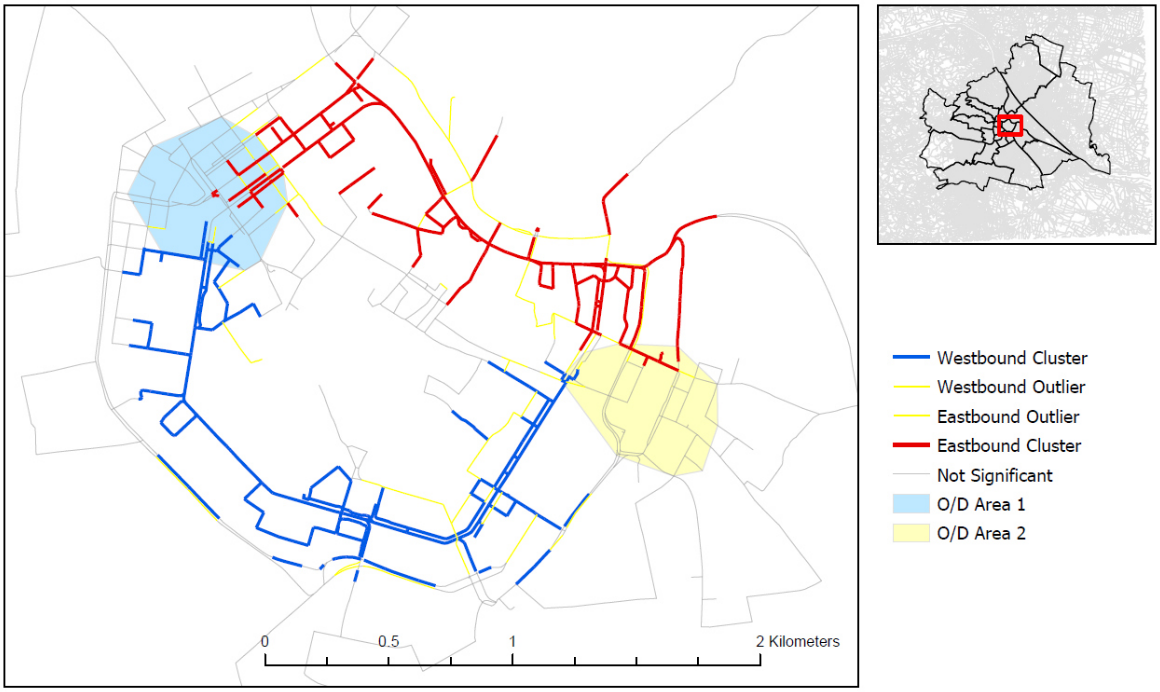

Figure 38: Spatial clustering results from Wien Mitte to Siegmund-Freud-Park and vice versa indicating areas of high mono-directional traffic flow (n=250).

Case Study 4 – UNO City to Airport

For the next case study, as in case study 3, two OD-areas were selected that lie far apart from each other. The straight-line distance between the airport in the south and UNO city in the north is approximately 15 kilometers. Therefore, there are potentially many viable routes to choose from. However, the results depicted in Figure 39 indicate that again most drivers seem to choose the same route. Furthermore, no directional effects are visible, drivers take the same route regardless of direction of travel. There are some northbound cluster along highway A4 – the route preferred by most drivers in this case study – however, there are also many outliers along this highway, so these mini-cluster do not mean that there are any clear directional effects.

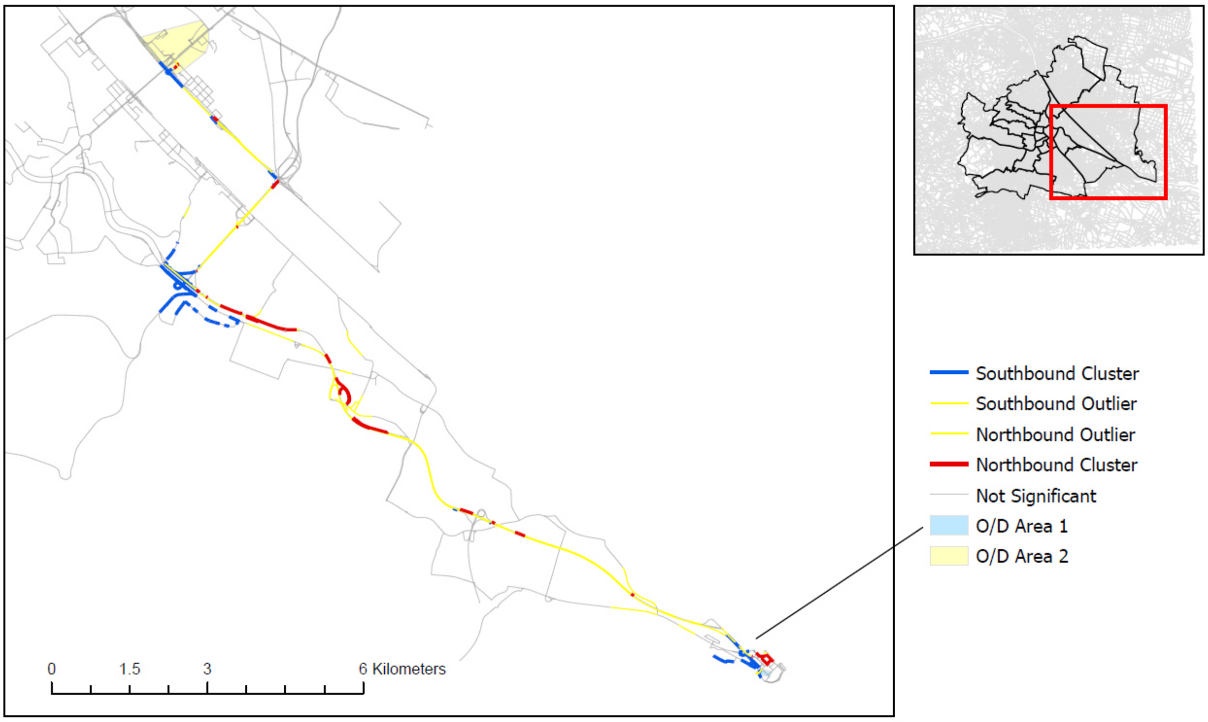

Figure 39: Spatial clustering results from UNO City to Schwechat Airport and vice versa indicating areas of high mono-directional traffic flow (n=500).

Case Study 5 – Wiener Westbahnhof to Wiener Staatsoper

Case study 5 involves trips from Wiener Westbahnhof (railway station) to the opera in the city center and vice versa. Judging from Figure 40, drivers mainly choose one route from east to west, namely along highway number 1, also known as Rechte Wienzeile. From west to east on the other hand, drivers tend to choose to follow Gumpendorfer Strasse to reach the opera. Again, the results are not a clear indication of any directional effects. Figure 41 illustrates the situation at the intersection of Gumpendorfer Strasse and Getreidemarkt. In this area, Gumpendorfer Strasse is a one-way street and therefore it is not possible to use it to go from the opera to the railway station in the west. Nevertheless, this is the first case study, where travel direction has some effect on route choice behavior of taxi drivers, as it would be possible to use highway 1 (Rechte or Linke Wienzeile) to go from east to west as well as from west to east. The most likely explanation for the fact that drivers prefer Gumpendorfer Strasse over highway 1 when going from west to east is that the route over Gumpendorfer Strasse is faster and drivers would also use this street in the opposite direction if it were possible.

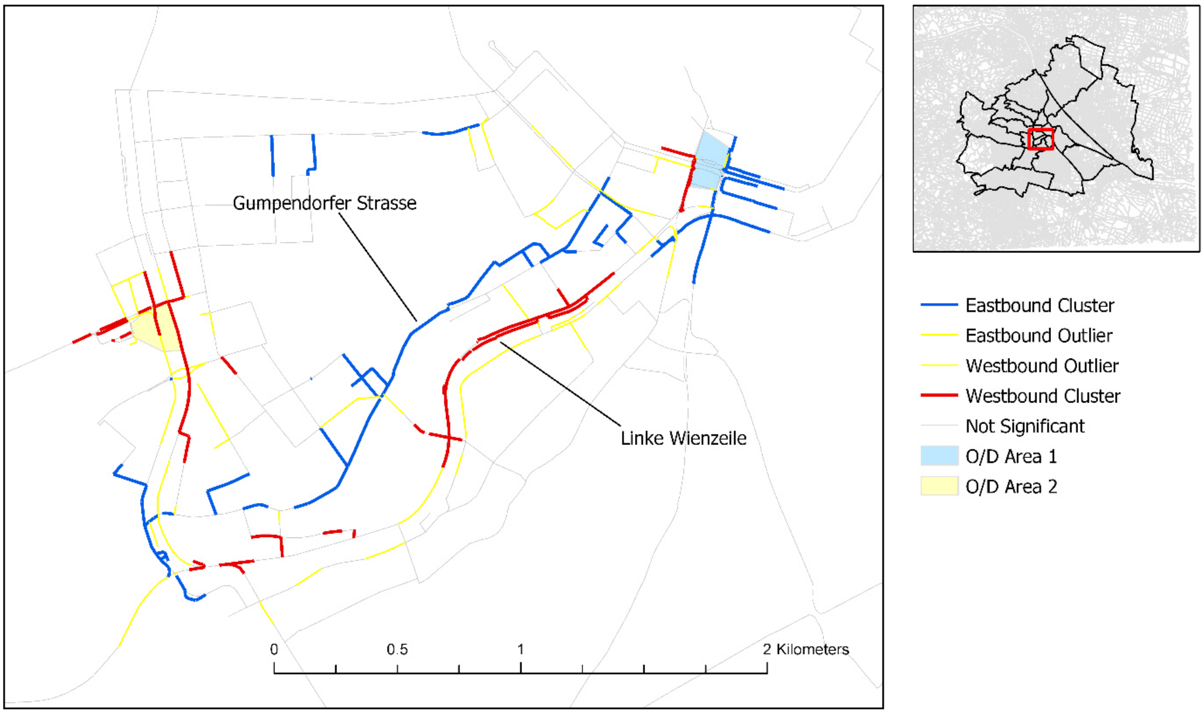

Figure 40: Spatial clustering results from Wiener Westbahnhof to Wiener Staatsoper and vice versa indicating areas of high mono-directional traffic flow (n=125).

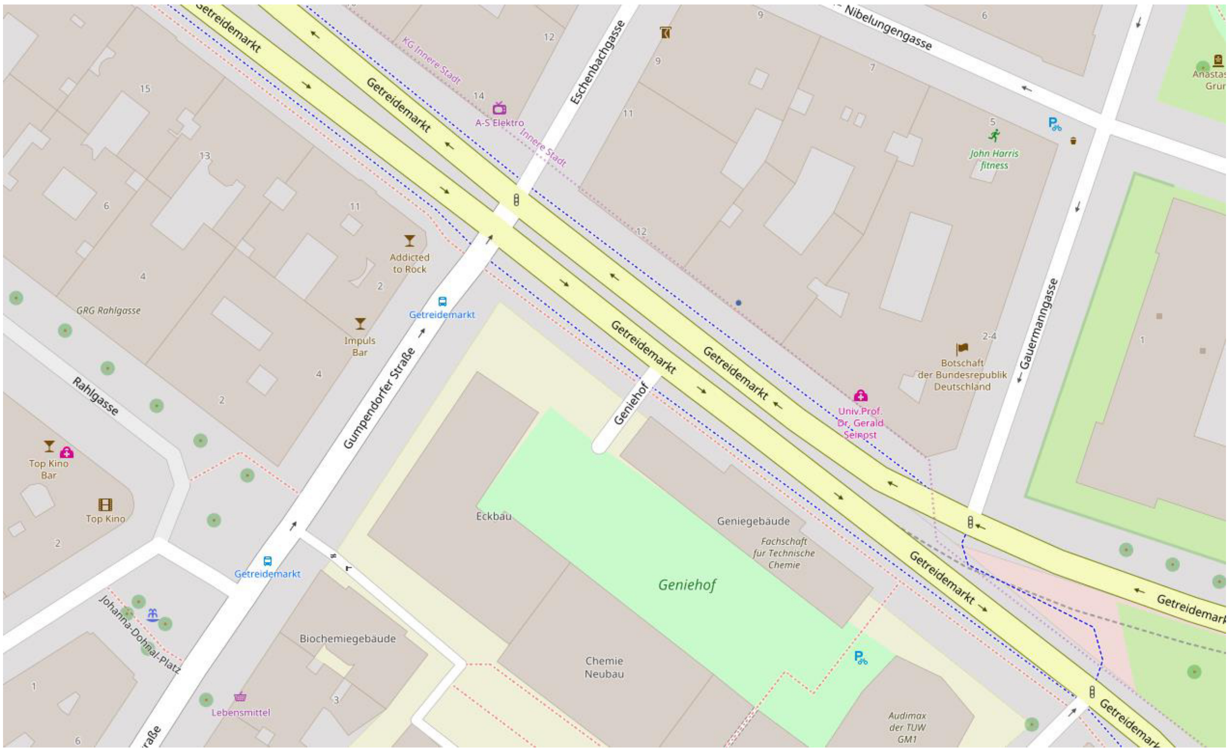

Figure 41: Situation at the intersection of Getreidemarkt and Gumpendorfer Strasse; map © OpenStreetMap contributors.

Case Study 6 – Florisdorf to Wieden

The final case study of this group consists of trips going from a district of Vienna – Wieden – located to the south of the city center to another district – Florisdorf – to the north of the city and vice versa. As the origin and destination areas are quite big this time, there are 1000 trips per direction in this case study, which is the largest number of all. There are again many potential routes to choose from. The clustering results can be seen in Figure 42. This is the first case study where there are clear directional effects which are not a result of street network constraints. Around the ring road we have the expected cluster due to one-way restrictions. To the north however, we can see a southbound cluster in the left along highway 227 and we can identify a northbound cluster east of the highway, where drivers use Jägerstrasse, Stromstrasse and Dresdner Strasse to approach Florisdorf. None of these streets are one-way streets. However, the directional effects are still not very distinct. For example, the path along the Danube is not affected by any directional effects. The cluster are not very far apart from each other. Therefore, it is not the case that taxi drivers choose entirely different routes based on direction of travel.

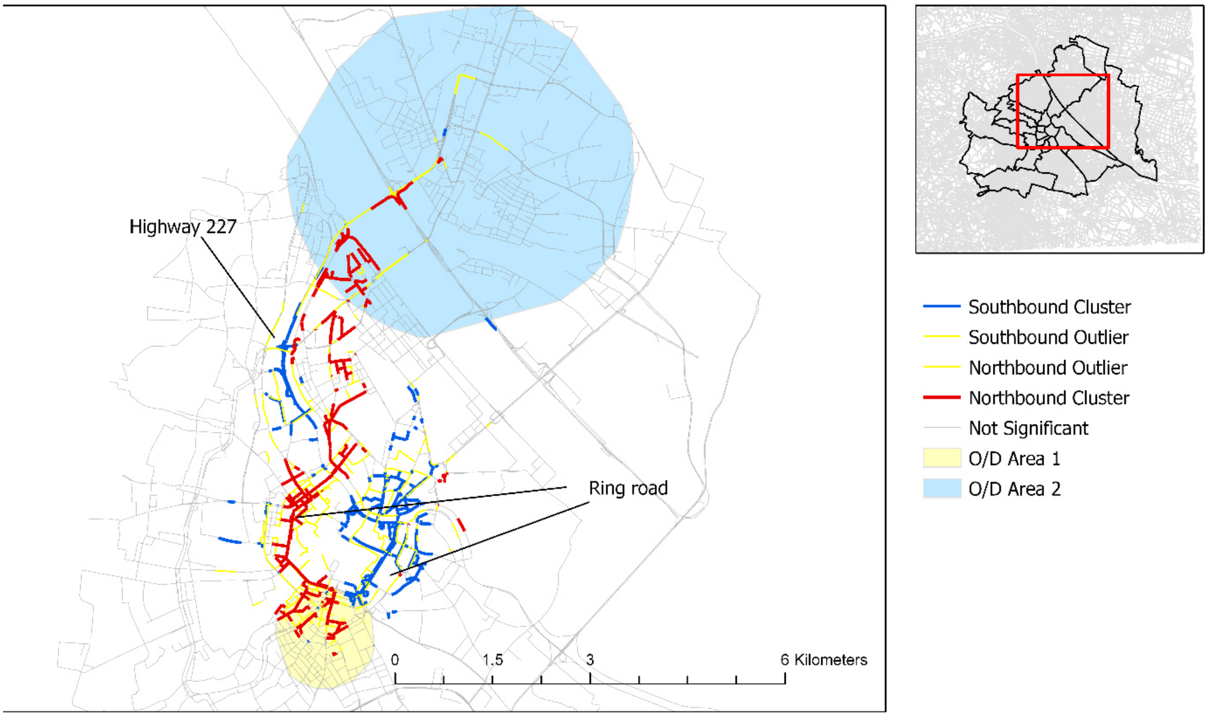

Figure 42: Spatial clustering results from Florisdorf to Wieden and vice versa indicating areas of high mono-directional traffic flow (n=1000).

To conclude, the case studies in this group have shown that there are very few, if any, directional effects that are not due to street network constraints. This means that taxi drivers in Vienna do not seem to be affected by a directional bias, in contrast for example to the minicab drivers from the Manley, Addison, and Cheng (2015) study. One possible conclusion from this is that Viennese taxi drivers take near optimal routes, as usually the fastest route from A to B is also the fastest to go from B to A. Furthermore, it could mean that taxi drivers do not plan their routes around anchors particularly, but rather have the complete route already in mind at the beginning of the trip.

5.3.4 Case study group 3: Centrality

The third group of case studies is a bit more diverse than the previous two, meaning there is more than one aspect that is investigated. However, all case studies in this group use one of the following three centrality indices for the analysis: betweenness, closeness and information. All these indices have been calculated globally for the whole street network of Vienna and its surroundings. Before we continue with the individual case studies, we shall look at these centrality indices of the street network.

The global betweenness centrality is depicted in Figure 43. Highway A22 along the northern bank of the Danube as well as highway 1, 8, 14, 227/A4, A21 and A23 have the highest values. Surprisingly, the ring road and highway 221 do not have very high values, their road segments are mainly classified as Medium-Low. From the five bridges than span over the Danube river that are accessible by car, only the most northern (Nordbrücke) and southernmost (Praterbrücke) bridges have very high betweenness values.

The aim of the first two case studies in this group is to check if there is any correlation between a high betweenness value and the popularity among taxi drivers of a road.

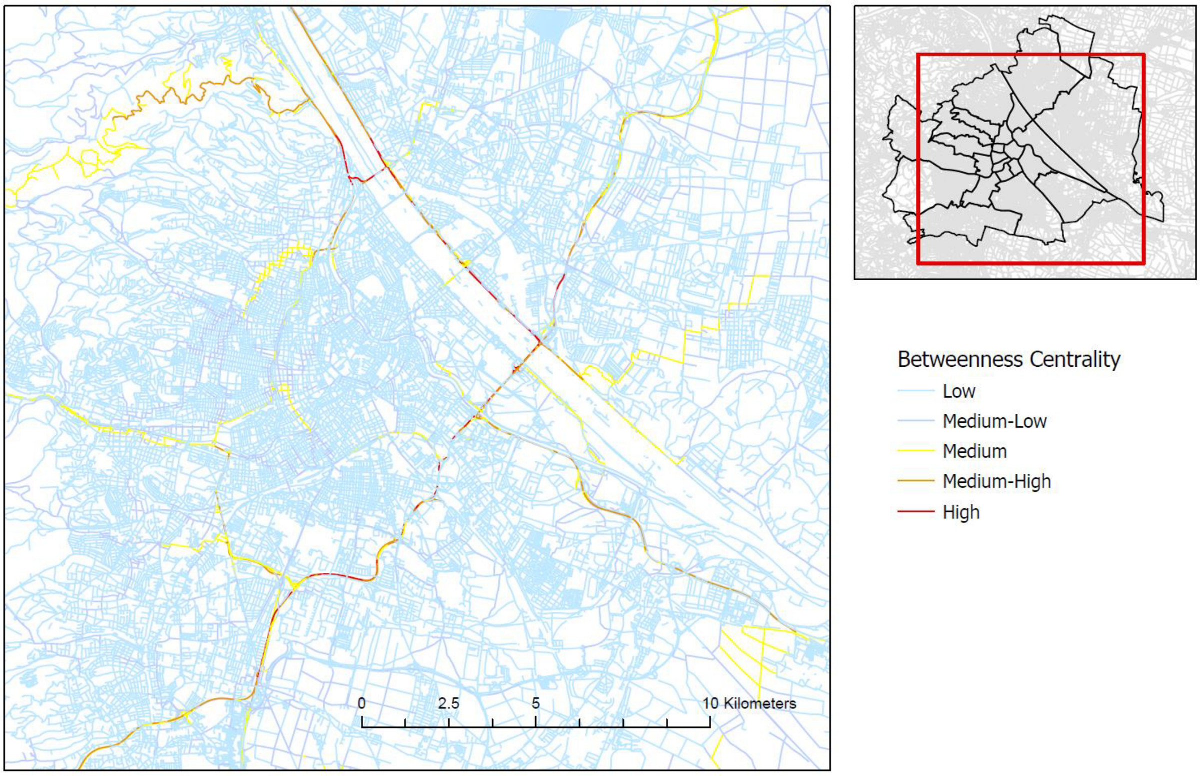

Figure 43: Global betweenness centrality of Vienna's street network.

Next, case studies 6, 7 and 8 involve the closeness centrality index, shown in Figure 44. The area left and right of the Danube river, the area along the highway A23 expose the highest closeness centrality values. All major bridges over the Danube have very high values. From these areas the closeness values decrease outwards to the periphery. However, there are some areas which present themselves as exemptions to this simplified pattern. The most prominent being the area around the historic city center. More specifically, parts of the districts of Neubau, Mariahilf and Innere Stadt have lower values than surrounding areas.

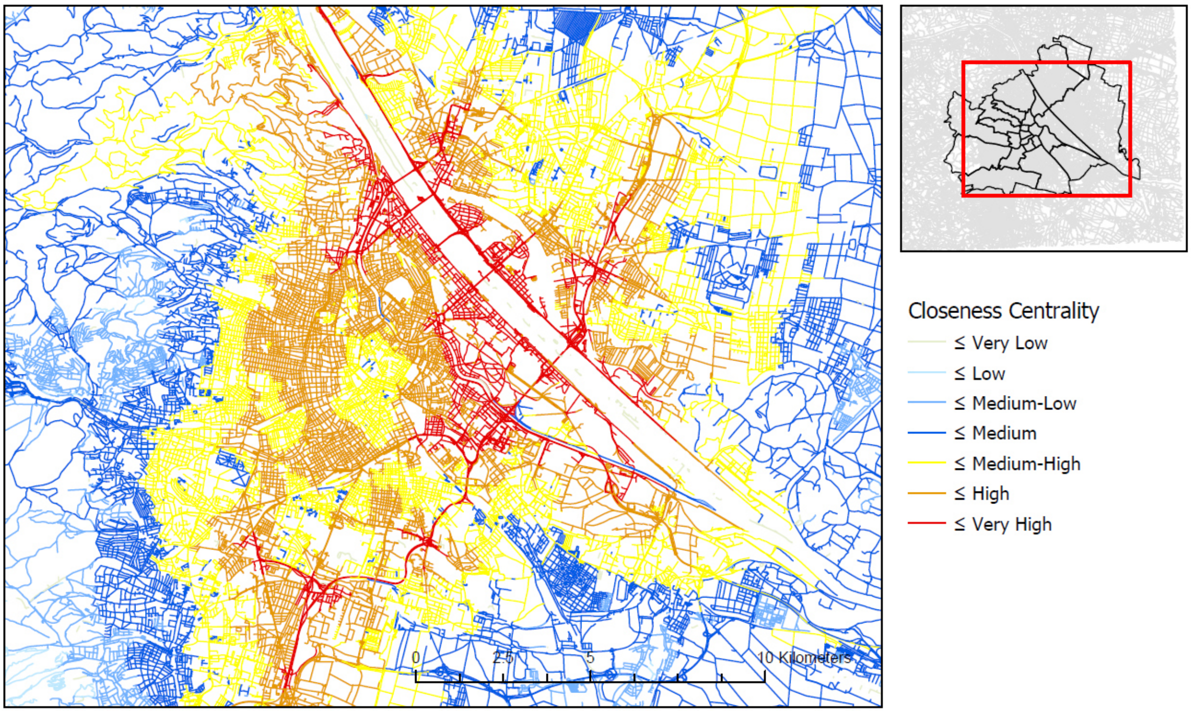

Figure 44: Global closeness centrality of Vienna's street network.

The last index, information centrality, is displayed in Figure 45. The pattern differs significantly from the one of Figure 44. The area around the Danube has relatively low information centrality values, especially along the northern bank of the river. The five bridges all have high values, however, only the two northern bridges (Nordbrücke and Florisdorfer Brücke) as well as the Reichsbrücke are put in the highest category. Generally, information centrality decreases from the city center towards the edge of the study region, but there are again some patches with very low values within areas with high values and vice versa. Here, it is again the inner city, which has low values not only compared to its surroundings, but in general. Information centrality is used in case studies 3, 4 and 5.

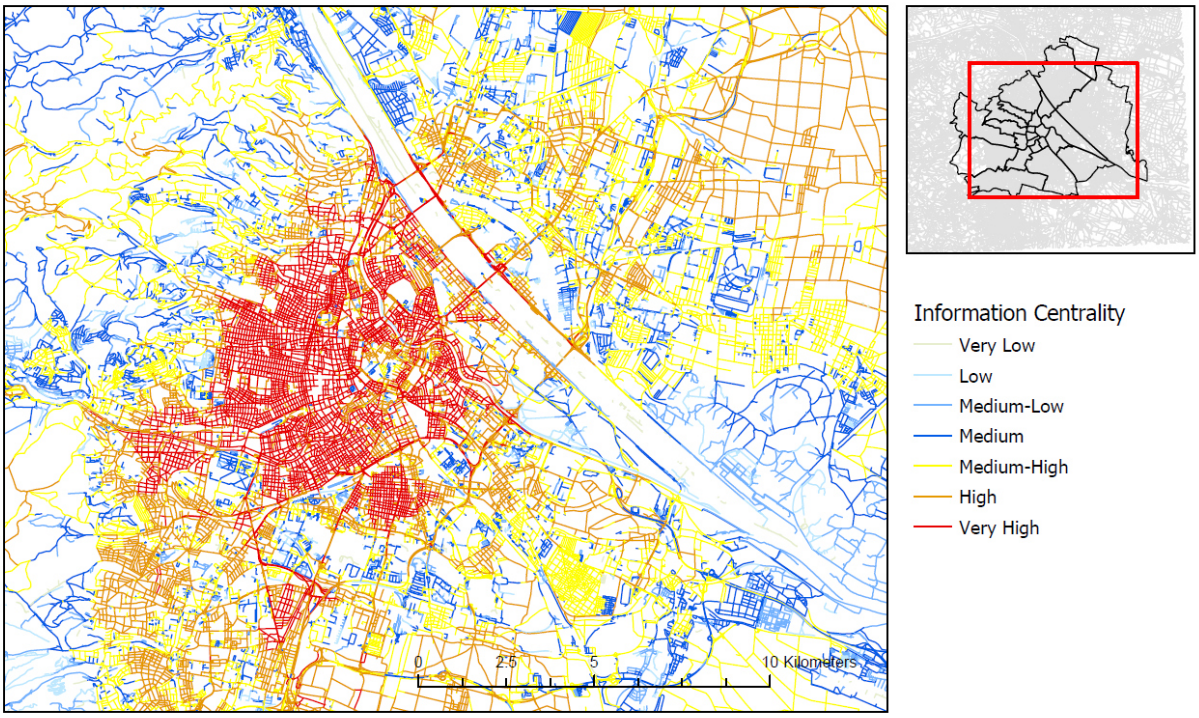

Figure 45: Global information centrality of Vienna's street network.

Case Study 1 – Correlation with betweenness centrality 1

The aim of the first case study in this group is to check the correlation of betweenness centrality and popularity of road segments. 250 trips have been selected for the analysis, all of which lead from the area around Am Spitz in Florisdorf in the north to an area south of the city center. If taxi drivers favored road segments with high betweenness centrality values, they would be expected to first drive on the A22 along the Danube and then going over the Praterbrücke, go to the roundabout Verteilerkreis Favoriten using highway A23 and then reach the destination area using highway 225. Figure 46 shows that taxi drivers indeed favor this route, with about two thirds driving on the roads described before. However, the remaining third of the drivers chose a different route. In this alternative route, drivers use A22 first, then go over then Danube on the Nordbrücke and lastly use highway 227 to reach the destination area. Looking at Figure 43 (same extent as Figure 46), one can see that this route has relatively low betweenness values (classified as Medium-Low) that do not stand out.

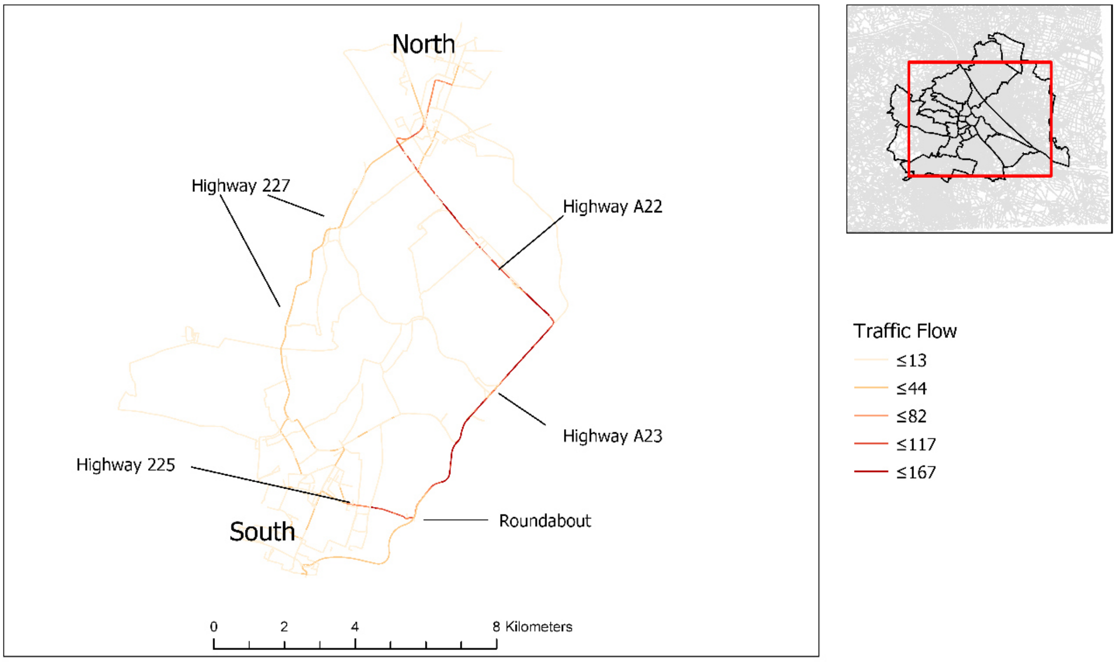

Figure 46: Traffic flow through Vienna from north to south (n=250).

Case Study 2 - Correlation with betweenness centrality 2

Figure 47 shows the traffic flows for case study 2. The origin area lies north of the city center, the destination area is Schwechat airport in the southeast of the city. Figure 48 shows global betweenness centrality values of the streets in the study area. If the hypothesis that taxi drivers prefer roads with high betweenness centrality were to be true, the expected routes would all follow highway A4 until Praterstern intersection and then continue on highway 227 until Friedensbrücke and finally reaching the destination area on Alserbachstrasse. This would be the path that maximizes betweenness centrality. Highway A4 has very high values, after Praterstern there are only roads classified as medium-low and low leading to the destination. The traffic flows from Figure 47 show that about half of the trips correspond to the maximum betweenness path. The other half of the trips is on highway 225 south of highway A4, then on Simmeringer Hauptstrasse and Rennweg before reaching the destination area by circumventing the city center on the left-hand side. It seems betweenness centrality alone is not a good predictor for observed traffic flows. Especially in the two case studies above, where origin and destination are far apart from each other, taxi drivers seem to prefer primary over secondary roads and secondary over tertiary roads.

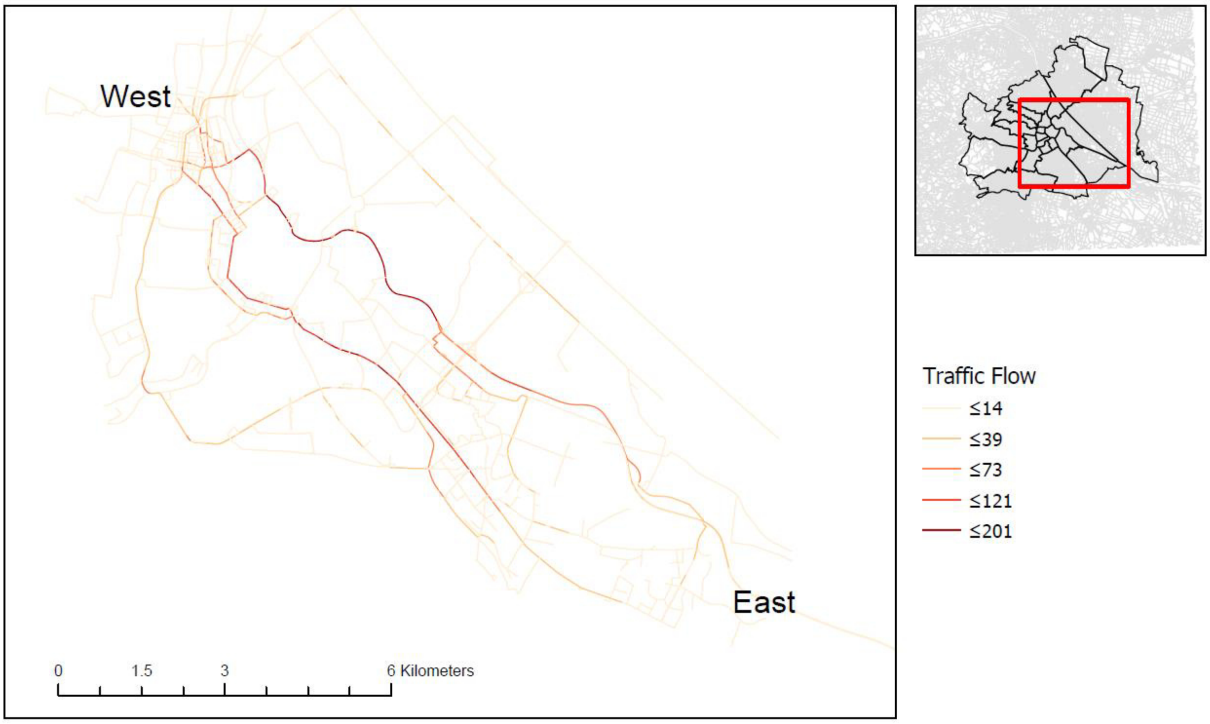

Figure 47: Traffic flow through Vienna from east to west (n=125).

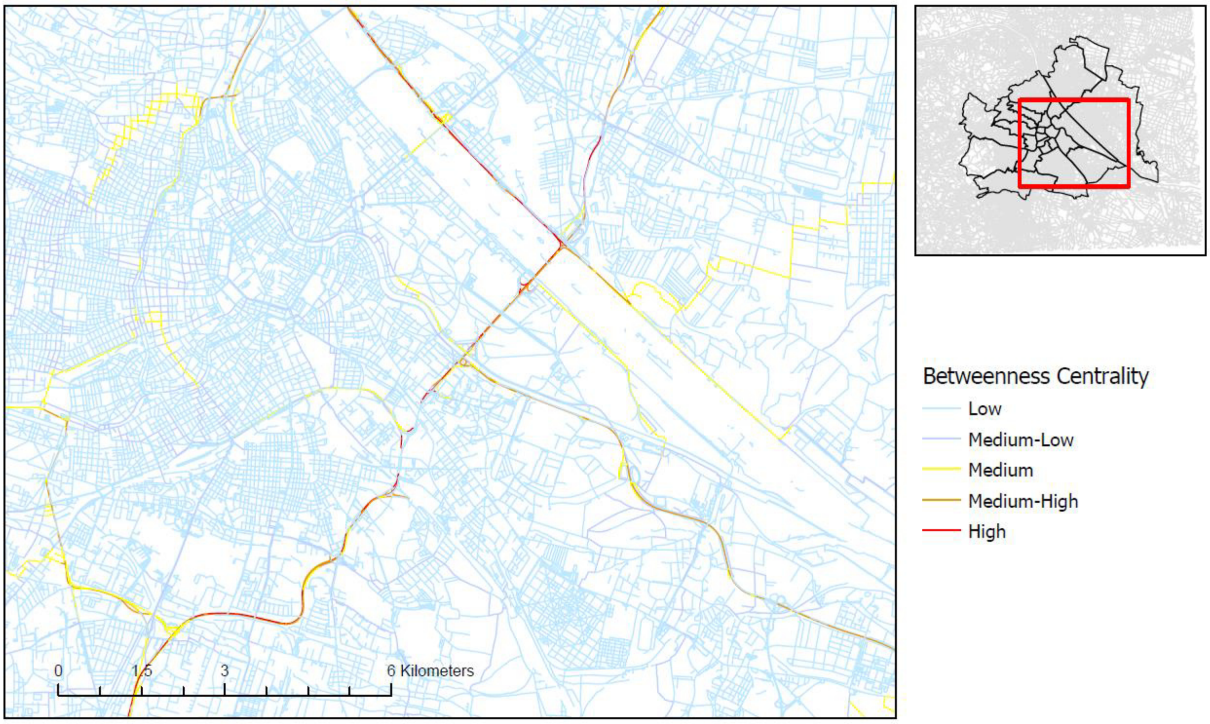

Figure 48: Global betweenness centrality of Vienna (same extent as in Figure 47).

Case Study 3 – Avoided areas – Information centrality I

In the next two case studies, the aim is to assess whether taxi drivers avoid or favor certain areas of a city based on differences in road network characteristics. More specifically, the aim is to check if taxi drivers prefer to drive through areas based with low information centrality or if they rather avoid such areas even in cases where this would result in longer travel distances. To do so, the area around the city center shown in the next two figures has been selected for the analysis. The traffic flow volumes from south to north are displayed in Figure 49. Figure 50 shows the corresponding street network colored according to the information centrality of the road segments. The area in the middle of the study region has significantly lower information centrality than the surrounding streets along the ring road. Since the direct route from south to north cuts straight through the city center, the analysis of the traffic flows in this scenario gives a strong clue about whether or not taxi drivers avoid (or prefer) areas with low information centrality. Looking at the traffic flows, the verdict is clear. The majority of drivers avoids going through the center and drives along the ring road. When examining the street network properties in the study area, one gets an idea why this might be. There are many turn and one-way restrictions in the area encircled by the ring road. Additionally, the speed limits are low in the center and higher for the ring road. Therefore, it is not surprising that most drivers choose the ring road, however, it is possible to go straight through the center and a minority of drivers actually did that. This analysis draws the conclusion that information centrality might capture the complexity of street networks (restrictions, such as one-way streets and turn limitations adding to the complexity). Additionally, this case study has shown that drivers clearly avoid streets with low information centrality values in this scenario.

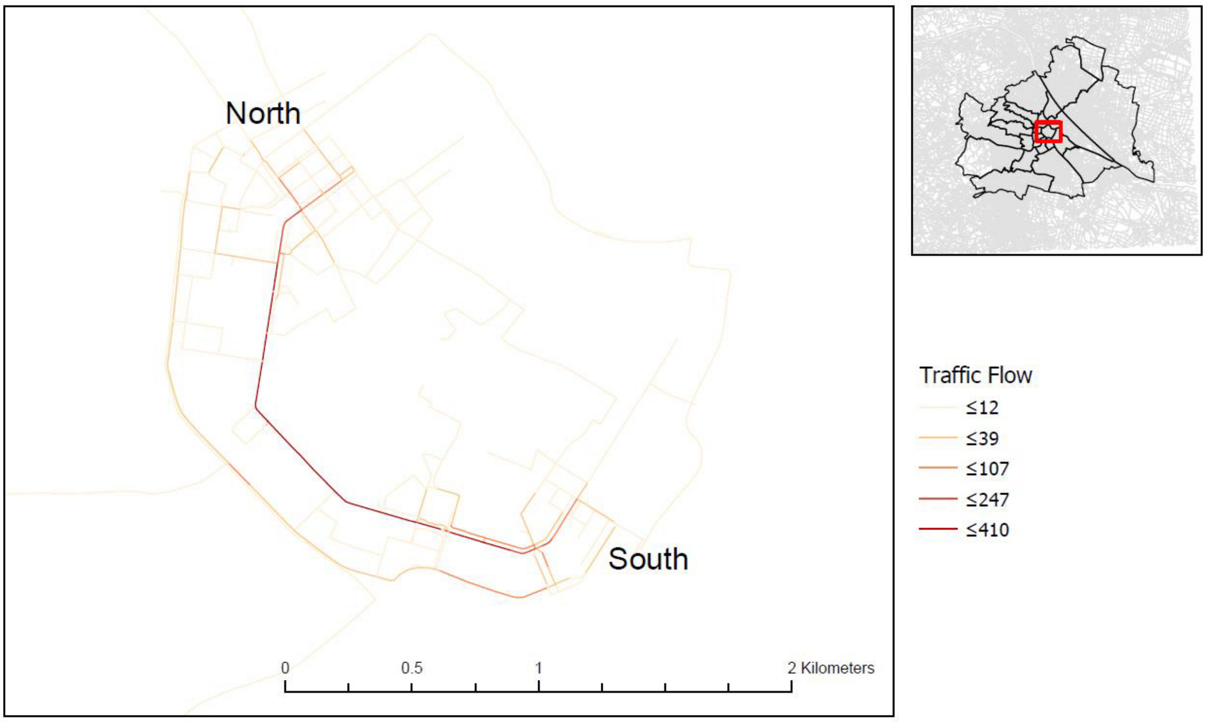

Figure 49: Traffic flow from Stadtgarten (south) to Siegmund-Freud-Park (north) (n=300).

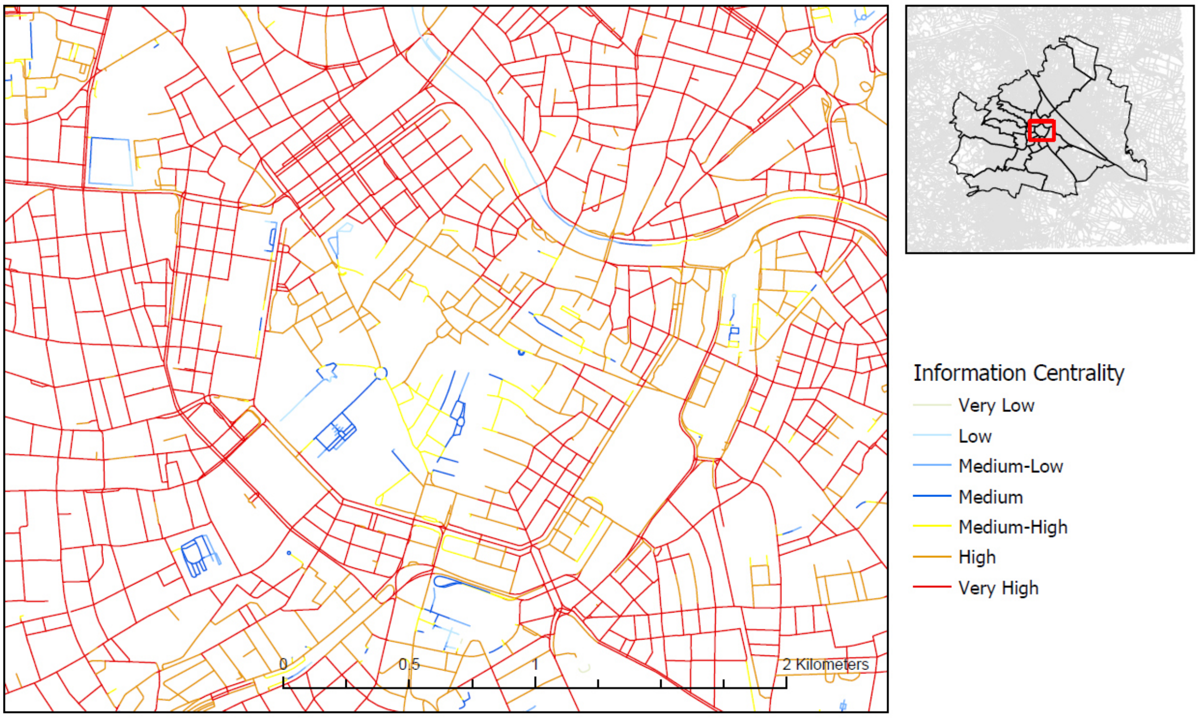

Figure 50: Global information centrality of Vienna (same extent as in Figure 49).

Case Study 4 – Avoided areas – Information centrality II

Case study 4 resembles case study 3 in that there is again a region in the center with relatively low information centrality values surrounded by streets with higher values including a highway that is clearly visible (see Figure 52). Hence, the objective is again to investigate if there is any visible preference for high or low information centrality roads among taxi drivers. The traffic flows (see Figure 51) reveal a pattern which contrasts the results from the previous case study. The maximum information path leads along the highway, but most observed paths cut straight through the case study area, minimizing travel distance. Only a small fraction of trips leads along the highway. These results have multiple implications. First, the resulting travel flows probably depend a lot on the choices for origin and destination. If there is a road with low information centrality and low speed limits but directly connecting origin and destination, most drivers seem to still use this road, at least that is the case in this scenario.

Still, it seems that road network density as a significant factor in terms of route choice. Cluttered areas or dense areas are sometimes ignored or traveled through quickly (Bailenson, Shum, and Uttal 1998). However, in this scenario, the results suggest, that the preference for straight routes has been more important.

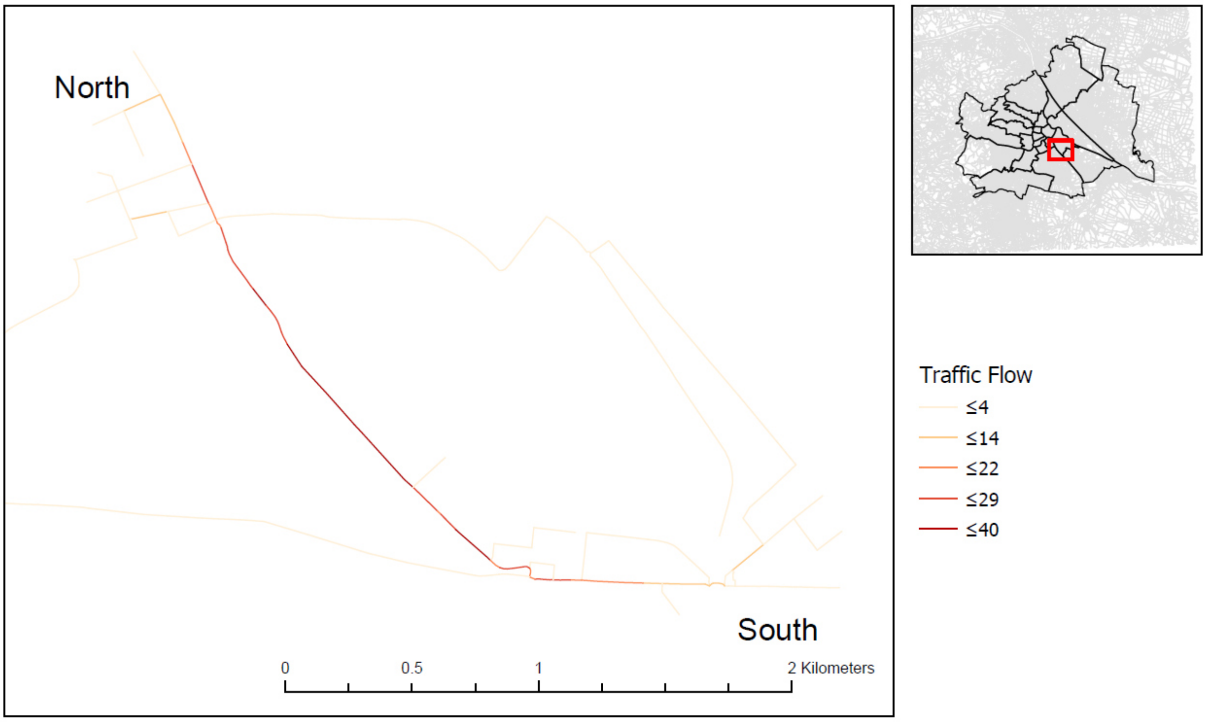

Figure 51: Traffic flow from Wien Geiselbergstrasse train station (north) to Belvedere Palace (south) (n=25).

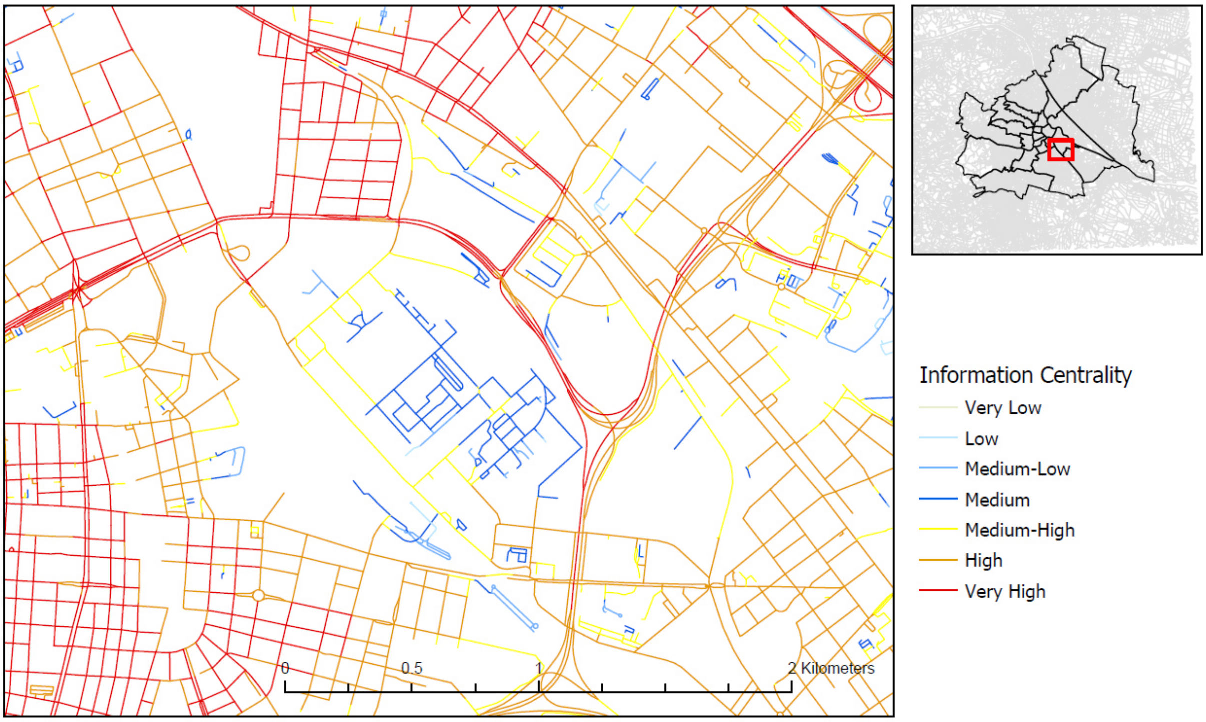

Figure 52: Global information centrality of Vienna (same extent as in Figure 51).

Case Study 5 – Comparison of trips in areas of high information centrality versus trips in areas of low information centrality

The next two case studies differ from the previous ones insofar as they do not incorporate a visual analysis of the trips themselves. Rather, the analysis is similar to the one in chapter 5.2 as it consists of a comparison of route characteristics of two groups of paths. However, here the trips are split based on location, and more specifically, based on the average information centrality in a region. Thus, the variable – centrality - that is investigated is dependent and therefore, it is included it in this part of the thesis and not in chapter 5.2.

All trips that start and end in the same of the two areas (see Figure 53) are identified. Subsequently, route characteristics have been calculated for 2000 trips per region.

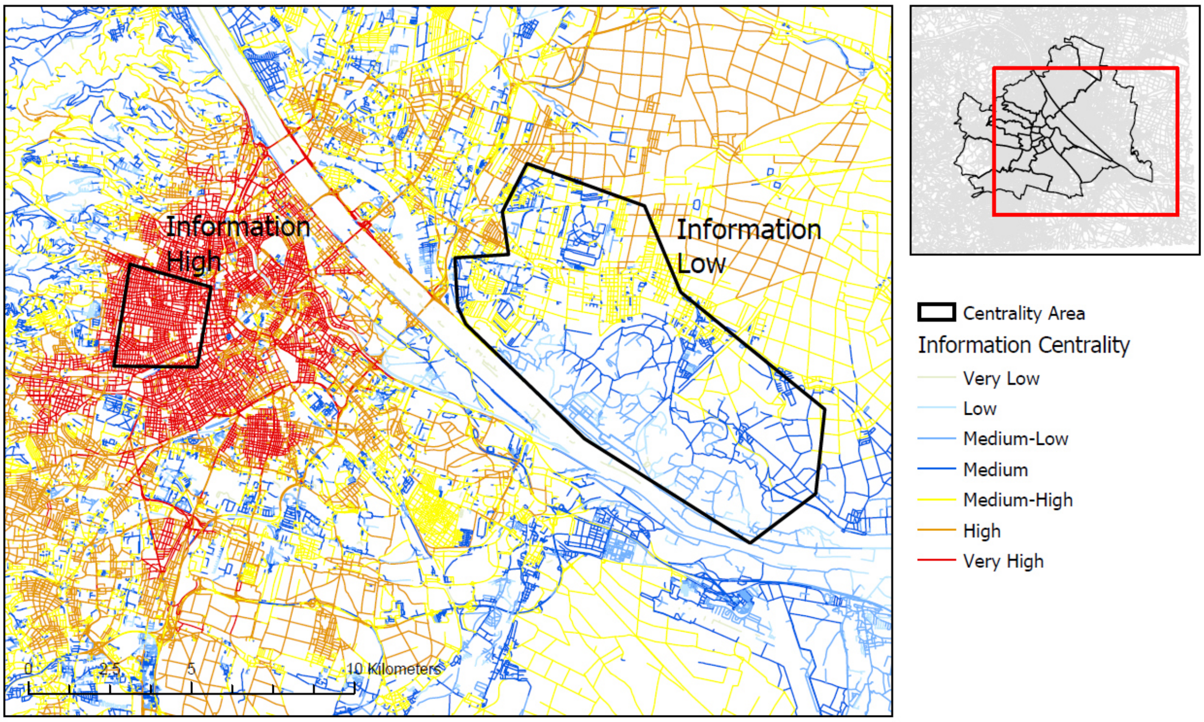

Figure 53: Map showing the location of the two groups of trips.

Since the region with low information centrality is significantly larger than the region with high information centrality, it is not surprising that the trips in the former group are also much longer on average than those in the latter group (see Table 19 for all results). Therefore, it is important to take into account these differences in route length when comparing the other attributes, as for example longer routes also tend to involve more turns and intersections. Keeping this in mind, it is evident that the trips in the high information centrality area involve significantly more intersections, which is not surprising, considering the much higher density of the network in this region. The number of turns is about the same when normalizing the results for route length. More interesting are the values for the RDI. The routes in the high value group are less direct than the ones in the other group, however, it is difficult to compare route directness in different study areas, as the configuration of the street network has a strong influence (Knight and Marshall 2015). The PSL values for the shortest paths are similar, but for the fastest paths, the overlap is higher in the case of low centrality values. This might come as a surprise, as there are many more alternatives to choose from in the case of high information centrality, which should theoretically lead to less overlap with optimal paths. As for the different centralities, it is evident that the high information centrality trips exhibit higher values for the other three indices as well. This could mean that the indices correlate positively with each other. The conclusion of this fifth case study is that the incorporation of space is important as it has a significant influence on route characteristics (Ciscal-Terry et al. 2016).

Table 19: Route characteristics of trips in areas with low and high information centrality values.

Variables	Low Information Centrality	High Information Centrality
Time (min) (free flow time)	4.35 ± 3.35	2.90 ± 4.93
Time (min) (Actual travel time)	6.30 ± 6.42	5.15 ± 6.60
Distance (m)	3087 ± 2481	1950 ± 1550
Number of intersections	0.99 ± 1.11	1.03 ± 1.74
Route directness index	1.76 ± 0.33	2.33 ± 0.40
Turns statistics		
Total all turns	3.46 ± 3.55	1.99 ± 2.70
% of route based on road type		
% distance on primary roads	0.29 ± 0.36	0.21 ± 0.32
% distance on secondary roads	0.08 ± 0.19	0.30 ± 0.34
% distance on tertiary roads	0.63 ± 0.39	0.49 ± 0.38
Overlap with corresponding optimal paths		
% of shared length with shortest paths	0.57 ± 0.39	0.57 ± 0.39
% of shared length with fastest paths	0.56 ± 0.40	0.54 ± 0.40
Centrality		
Betweenness centrality	0.00114 ± 0.00148	0.00300 ± 0.00283
Closeness centrality	0.00779 ± 0.00029	0.00832 ± 0.00014
Information centrality	0.8674-06 ± 4.0786-08	1.1670-06 ± 2.5981-08
Degree centrality	2.9248-05 ± 2.5776-05	3.3036-05 ± 3.3158-05

Case Study 6 – Avoided areas – Closeness centrality I

Case studies 6 and 7 are similar to case studies 3 and 4, only this time the focus is on closeness centrality instead of information centrality. The study area depicted in Figure 55 is located in the center of the city, where the area in yellow has lower closeness centrality values than the other areas in orange and red. The aim here is to see if closeness centrality has an influence on the route choice behavior of taxi drivers. Looking at the traffic flow in Figure 54 and comparing them with Figure 55 leads to the conclusion that taxi drivers do not seem to avoid areas with low values, as most trips cut through the yellow area. Considering that the study area here is the same as in case study 3, the results can mean that drivers either avoid roads and areas with low information centrality or that they prefer roads with low closeness centrality (a third possibility would be that both conclusions are true).

Figure 54: Traffic flow from Schweizergarten (south) to Nussdorfer/Alserbachstrasse (north) (n=50).

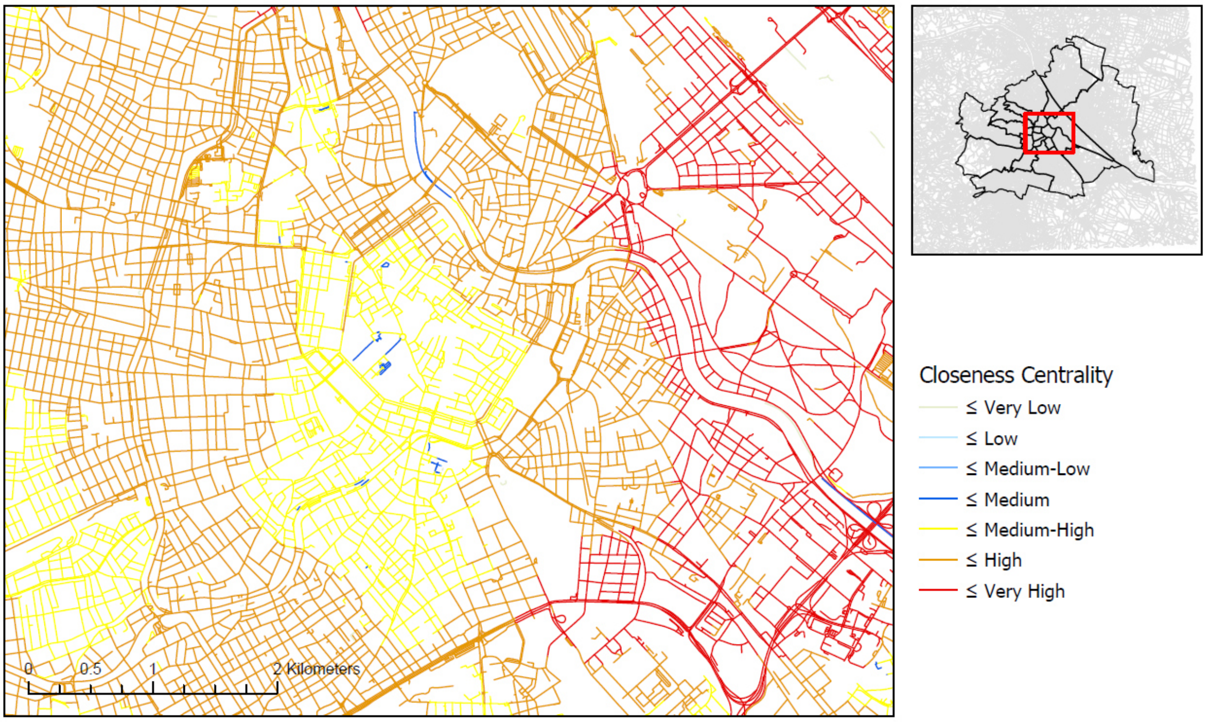

Figure 55: Global closeness centrality of Vienna (same extent as in Figure 54).

Case Study 7 – Avoided areas – Closeness centrality II

The origin area of this case study is around Kagraner Brücke in the east. The destination area is around Am Spitz in Florisdorf in the west (Figure 56). There are many options to choose from to go from origin to destination. The southern routes along the Danube have higher values than the northern routes (see Figure 57). The observed paths are mostly overlapping with the latter routes having lower closeness centrality on average. Thus, in this scenario, taxi drivers do also not tend to prefer roads with high closeness centrality values. However, it is important to note that the result depends a lot on the locations of origin and destination. In this scenario, the majority of trips starts north of Kagraner Brücke, which presumably affects the chosen paths a lot in that the southern routes would involve detours that are too big, so they are not in the consideration set of the drivers.

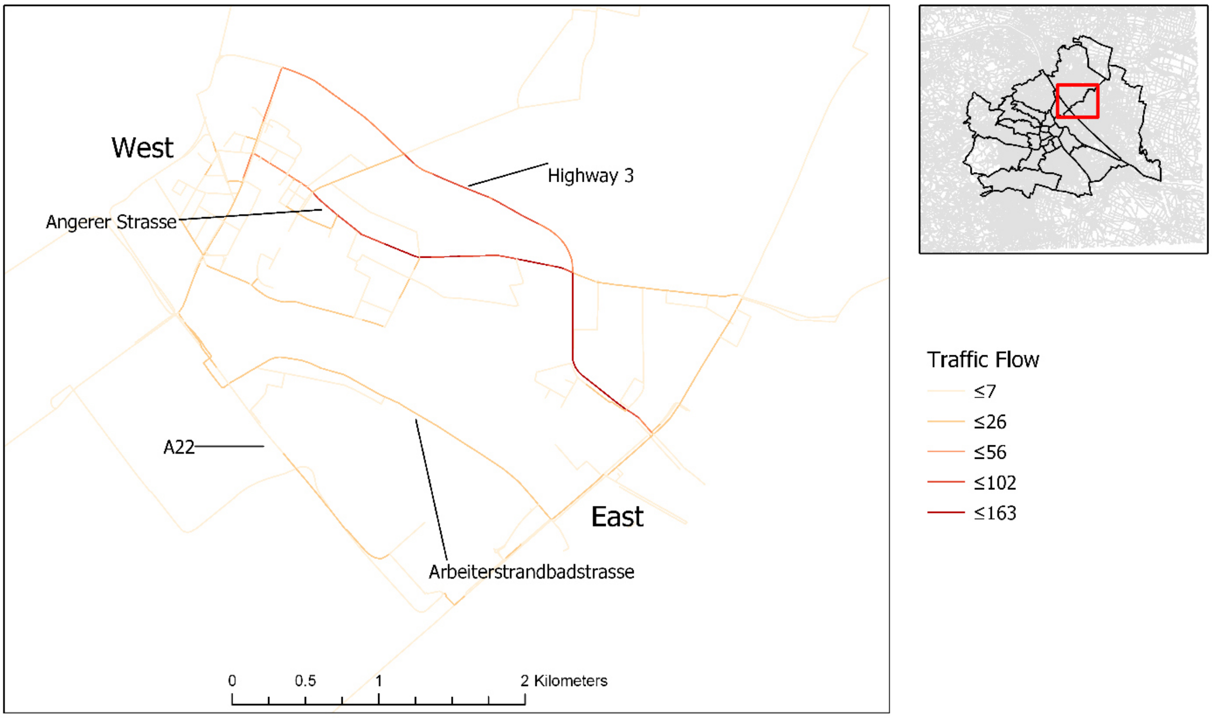

Figure 56: Traffic flow from Kagraner Brücke (east) to Am Spitz (west) (n=150).

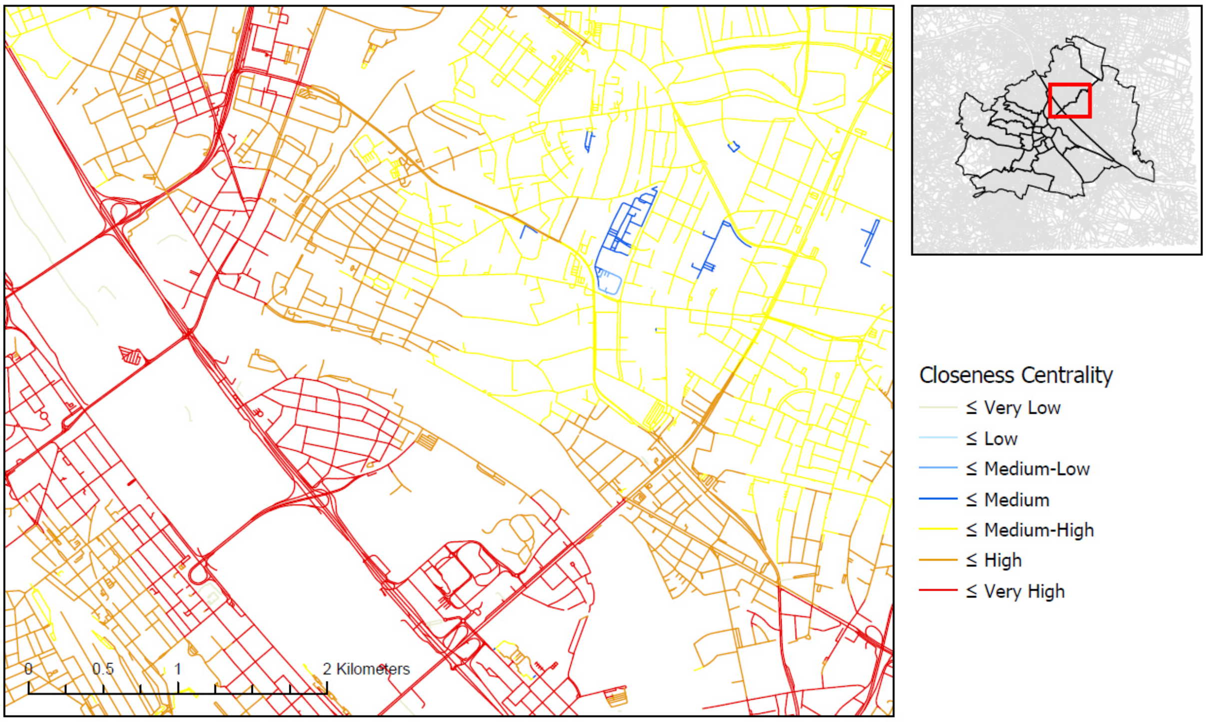

Figure 57: Global closeness centrality of Vienna (same extent as in Figure 56).

Case Study 8 – Comparison of trips in areas of high closeness centrality versus trips in areas of low closeness centrality

The procedure for case study 8 is similar to the one for case study 5, again the difference is the centrality index that is investigated, which in this case is closeness. Again, it is important to incorporate the fact that the trips in the high closeness group are about ¼ bigger than those in the other group.

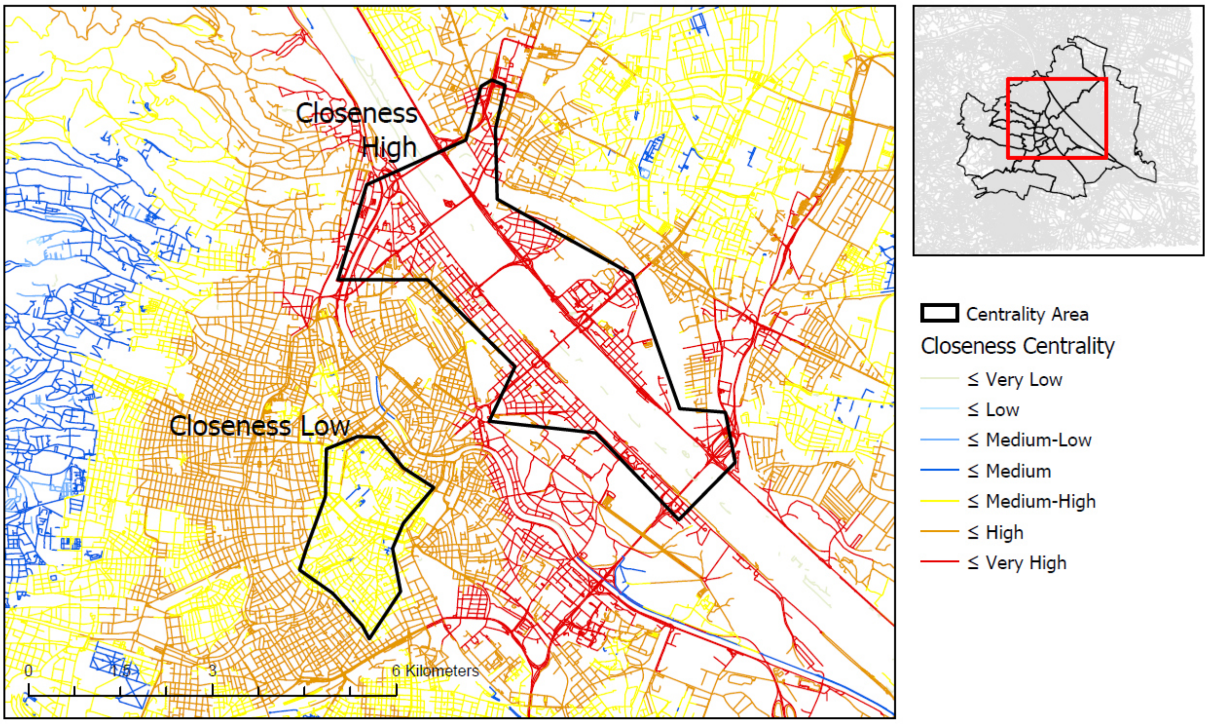

Figure 58: Map showing the location of the two groups of trips.

The resulting route characteristics are shown in Table 20. In general, the differences between the two groups are smaller than in case study 5. The trips in the low value group have more turns, the amount of intersections and the RDI is about the same for both groups. There are differences in the distribution of the different road types, but this is merely a reflection of the distribution of the relative frequency of the road types in the study areas. The most interesting characteristics to look at is the percentage of shared length. The PSL for both groups is much higher than the PSL values of chapter 5.1, 5.2 and those of case study 5. Additionally, the PSL values are significantly higher in the high closeness centrality group.

Table 20: Route characteristics of trips in areas with low and high closeness centrality values.

Variables	Low Closeness Centrality	High Closeness Centrality
Time (min) (free flow time)	2.35 ± 2.52	2.45 ± 3.00
Time (min) (Actual travel time)	6.03 ± 9.08	5.59 ± 7.56
Distance (m)	1512 ± 1780	1970 ± 1256
Number of intersections	1.25 ± 1.96	1.38 ± 1.99
Route directness index	2.09 ± 0.63	1.96 ± 0.43
Turns statistics		
Total all turns	3.89 ± 4.21	3.11 ± 3.94
% of route based on road type		
% distance on primary roads	0.21 ± 0.29	0.30 ± 0.37
% distance on secondary roads	0.20 ± 0.30	0.15 ± 0.26
% distance on tertiary roads	0.59 ± 0.38	0.55 ± 0.40
Overlap with corresponding optimal paths		
% of shared length with shortest paths	0.59 ± 0.41	0.65 ± 0.39
% of shared length with fastest paths	0.56 ± 0.42	0.62 ± 0.39
Centrality		
Betweenness centrality	0.00091 ± 0.00095	0.00283 ± 0.01043
Closeness centrality	0.00793 ± 0.00013	0.00930 ± 0.00031
Information centrality	1.0491-06 ± 5.9733-08	1.0349-06 ± 5.5405-08
Degree centrality	2.8013-05 ± 2.1463-05	2.9281-05 ± 2.2748-05

The analyses from this last group of case studies has not revealed any clear influences of centrality indices on route choice behavior. Betweenness centrality captures the importance of streets as other studies have also shown, but this index alone is not enough to predict chosen routes. As for the avoidance of certain areas, information and closeness centrality could only to some extent highlight such areas. It seems that information centrality captures complexity in a way (areas with a lot of one-way streets and turn restrictions receive a lower centrality values). These areas are usually avoided by taxi drivers. However, for all of this, more case studies are needed to come up with new hypotheses and to prove or discard existing ones.

6. Discussion

The analyses outlined in this thesis have described the chosen routes of taxi drivers in Vienna and assessed how time of the day, day of the week and weather conditions affect the route characteristics. Furthermore, the route choice behavior and the effect of road network centrality has been examined with detailed case studies. In this section, the findings are summarized and discussed, before they will be positioned within the larger context of route choice behavior research. Following this, it is discussed what real-world applications the findings could have and what the implications are. Last, some possibilities for future research in this area is outlined in the next chapter.

Route characteristics and overlap with optimal paths

The first part of the analysis has revealed that observed, shortest and fastest paths differ significantly from each other in term of route characteristics. Observed paths are less direct than both shortest and fastest paths with the observed paths being about 60 percent longer than the straight-line distance between origin and destination, compared to 25 or 30 percent, respectively, for shortest and fastest paths. Observed paths also have significantly fewer turns per kilometer than computed paths, indicating, that drivers prefer to minimize the number of turns and prefer to drive on straight streets as long as possible, at the expense of more direct routes. For some variables, the observed paths share more similarities with shortest paths, while the opposite is true for others. Generally, the overlap of observed and computed paths is only a bit more than 50 percent. This confirms findings from other studies, which have shown that minimum distance and travel time paths are bad predictors of observed paths (Manley, Addison, and Cheng 2015).

Influence of external factors

Next, it has been confirmed that time of the day and – to a lesser extent – day of the week have a strong influence on the chosen routes of taxi drivers. During times and days with presumably more traffic, taxi drivers seem to switch to tertiary roads at the expense of primary roads, which is probably faster since primary roads are more congested despite the fact that tertiary roads having lower speed limits. Additionally, the overlap of observed with computed paths increases the more traffic there is. The influence of the weather remains unclear, there are no differences in terms of route characteristics between different weather conditions that are significant and relevant.

Behavior in case studies

One important note to take away from the case studies is that the street network configuration varies strongly in space and thus the route characteristics and probably also the route choice behavior is very much dependent on location. Additionally, we have seen that while route characteristics over groups of paths does not always vary that much, this does not necessarily mean, that taxi drivers choose the same routes under varying conditions. There were clusters in most case studies for all three external factors, even for weather conditions. Here, it remains unclear from what these cluster exactly stem from, but there is a strong indication that external factors have a considerable influence on the route choice behavior of taxi drivers. In the case of the factors time of the day and day the of week this could be expected. Relative to other drivers, taxi drivers are more sensitive to travel time and more inclined to choose a route with less travel time (Hong-cheng, Xin, and Qing 2010). Therefore, they usually try to minimize travel time and thus pay close attention to the traffic conditions and choose different routes based on the amount of traffic. On the other hand, the influence of changing weather conditions is still unclear and further investigation is needed to be able to get to any meaningful conclusions.

In relation to directional effects, it became visible that all directional effects can be explained with the configuration and the restrictions of the road network. There were seldom any clear directional effects apart from the area around the ring road in the center of the city, which is a one-way street.

Influence of Centrality

In terms of the overlap with road segments with high centrality values, case studies revealed that there exist other factors than high betweenness centrality values that attract the taxi drivers and influences their behavior. Some highways and other routes have high betweenness values and those seem indeed to be popular among taxi drivers. However, they are also using a lot of routes that have low betweenness values, even in cases were there would be viable alternative routes with higher centrality values. To conclude, betweenness centrality can reveal the relative importance of streets (Tomko, Winter, and Claramunt 2008), but there are other factors that determine which routes are chosen. Especially the road types play an at least equally significant role, as we have seen that drivers switch routes depending on time of the day and day of the week. As betweenness centrality values remain constant over time, they cannot be the single most crucial factor determining the route choice.

Avoided and preferred areas

Another aim of this study was to try to come up with a measurement for complexity of areas and to check, whether taxi drivers avoid such areas or not. In this study, closeness and information centrality have been chosen as measurements for this. The case study analyses could not reveal any definitive patterns in the case of closeness centrality.

Nevertheless, especially information centrality lead to some promising results. In the case studies of the present work, information centrality proved to be a good indicator of the attractiveness of road segments. Areas with a lot of turn and one-way restrictions, as well as areas that are weakly connected to other areas of the street network, generally exhibit low information centrality values. Taxi drivers at least in one case study do not travel through such areas very often. Therefore, calculating information centrality could be an uncomplicated way to capture areas with a lot of turn and one-way restrictions, which are probably relatively unpopular among drivers. However, more case studies are needed to be able to confirm this theory.

Route choice behavior of taxi drivers – Anchor-based or shortest-path route choice?

Following Manley et al. (2015), in this study it was investigated if either the shortest-path model or the anchor-based model can explain the route choice behavior of taxi drivers. As discussed earlier, the shortest-path model is most likely not suited for this task. The shortest paths in this study (shortest distance and minimal free-flow travel time paths) do not highly overlap with observed paths. Additionally, betweenness or any other centrality index alone is not suited to predict chosen routes adequately and any potential connection must be investigated further.

Directional effects were almost non-existent in the case studies. In cases where there were some directional effects, they were always a consequence of street network properties, which indicates that taxi drivers do not seem to plan their routes around anchors. However, as the identification of anchors themselves has not been part of the present analyses, it is not possible to falsify the anchor-based model completely for Viennese taxi drivers.

What drives the behavior then?

If both above models seem to be unsatisfactory for the present scenario, the question is: What drives the behavior of the drivers? To answer this, first we need to look at the fact that the drivers' behavior is rather homogenous. In all instances where trips from A to B have been investigated, the majority of drivers took the same or highly similar routes. Furthermore, taxi drivers seem to drive optimal routes At least one can say that directional effects are difficult to find, which is a strong indication that they do not plan their paths *en route*, making decisions at specific important points or landmarks, but that they plan their routes beforehand. Maybe they also adapt spontaneously to changing traffic conditions. The key point both for the homogeneity and the optimal route choice, could be the experience. Licensed taxi drivers need to learn quite a lot about the street network of Vienna and as they are professional drivers, they spend a lot of time driving around the city, which enables them to find optimal routes over time (Stern and Leiser 1988). Relative to other drivers, the ones with a rich driving experience are less likely to choose a route with travel time uncertainty (Hong-cheng, Xin, and Qing 2010). This is reflected in the results of the present analysis. Consequently, this means, that the anchor-based route choice model from Manley et al. is not applicable entirely to taxi drivers in Vienna as they do not seem to show any differences in the behavior depending on direction of travel. However, it is not possible to draw any final conclusions, as the floating car data of this study does not contain any information about the drivers and therefore, the role of experience cannot be analyzed directly.

7. Conclusion

The analyses outlined in this thesis have shown that taxi drivers usually do not follow shortest distance paths nor minimum travel time paths. Observed and both groups of computed paths have significantly different route characteristics. Thus, as many other studies have already shown for other groups of drivers (Papinski and Scott 2011; Manley, Addison, and Cheng 2015; Sun et al. 2014), the shortest path model does not sufficiently represent the behavior of drivers.

However, it is unclear if the anchor-based model performs better. While the identification of anchors has not been part of the analyses, the results from the exploration of directional effects question the validity of the anchor-based model for the case of taxi drivers to some extent. In the scenarios that have been looked at in this study, taxi drivers did not show any clear directional biases. Additionally, there is evidence that taxi drivers in Vienna are rather homogenous and choose similar routes for similar OD-pairs. Assuming that there is no wide-spread use of in-car navigation systems, these findings indicate that taxi drivers do not plan their routes around anchors necessarily. If every intersection becomes an anchor, the model is not very helpful. Further research is needed to see if taxi drivers have their routes in mind beforehand, and whether these routes are really near-optimal in terms of travel time.

Also, chosen paths differ significantly in terms of route characteristics depending on the time of the day. Case studies revealed that during rush-hour taxi drivers choose different routes than during times with less traffic. They do this most likely to avoid traffic and minimize travel times.

For future research it would be interesting to create a comprehensive framework which includes findings of this study to predict routes chosen by taxi drivers to enhance traffic models. Furthermore, some of the analysis methods from this and other studies have not yet been applied to ordinary drivers to the best of my knowledge. Especially the spatial clustering method could reveal interesting results when analyzing directional and external effects of ordinary drivers. As the majority of drivers on the roads are not professional ones, such an analysis would be highly relevant. For many analyses in this thesis, it is necessary to conduct more case studies in other scenarios, to be able to verify or falsify the findings. This is especially true for the case studies where a connection between popular and unpopular areas and information or closeness centrality has been investigated. Additionally, more case studies are needed to assess the influence of weather conditions more accurately. Future research should hold constant other factors, to isolate the effect of changing weather on route choice behavior, which leads to more robust conclusions.

Last, I would like to give an answer to the question in the title of this thesis: No, taxi drivers do not take the shortest distance paths. However, there is robust evidence in the data that they take near optimal paths in terms of travel time in a lot of cases. There are three main points that lead to this assumption: First, drivers do not show any clear directional bias. This could be a sign of near-optimal route choices as one would expect that the fastest path in one direction is also fastest when going the opposite way. Second, overlap between observed paths connecting similar OD-areas is quite high. This leads to the conclusion that taxi drivers are on a similar level in terms of spatial knowledge (assuming they do not follow routes suggest by navigation systems that is). More importantly, as they are professional drivers one can assume that these paths are optimal in terms of travel time. The third point reinforces this assumption: Cases studies have revealed that taxi drivers choose distinct paths depending on the time of the day, where again, observed paths for a specific period are highly overlapping. This could well be a sign for a high sensitivity for travel time among the taxi drivers. The fastest path from an origin to a destination is not always the same, but changes depending on the time of the day since the amount of traffic varies heavily over the course of a day. For these reasons, the exploitation of taxi drivers' experience for the enhancement of navigation systems as well as for the improvement of traffic models, is very promising and should be researched further.

Literature

Akin, Darcin, Virginia P. Sisiopiku, and Alexander Skabardonis. 2011. “Impacts of Weather on Traffic Flow Characteristics of Urban Freeways in Istanbul.” *Procedia - Social and Behavioral Sciences*, 6th International Symposium on Highway Capacity and Quality of Service, 16 (January): 89–99.

Bailenson, Jeremy N., Michael S. Shum, and David H. Uttal. 1998. “Road Climbing: Principles Governing Asymmetric Route Choices on Maps.” *Journal of Environmental Psychology* 18 (3): 251–64.

Chen, Yang, Arturo Ardila-Gomez, and Gladys Frame. 2017. “Achieving Energy Savings by Intelligent Transportation Systems Investments in the Context of Smart Cities.” *Transportation Research Part D: Transport and Environment* 54 (July): 381–96.

Ciscal-Terry, Wilner, Mauro Dell’Amico, Natalia Selini Hadjidimitriou, and Manuel Iori. 2016. “An Analysis of Drivers Route Choice Behaviour Using GPS Data and Optimal Alternatives.” *Journal of Transport Geography* 51 (February): 119–29.

Datla, Sandeep, and Satish Sharma. 2010. “Variation of Impact of Cold Temperature and Snowfall and Their Interaction on Traffic Volume.” *Transportation Research Record: Journal of the Transportation Research Board* 2169 (December): 107–15.

Edwards, Julia B. 1999. “Speed Adjustment of Motorway Commuter Traffic to Inclement Weather.” *Transportation Research Part F: Traffic Psychology and Behaviour* 2 (1): 1–14.

Freeman, Linton C. 1977. “A Set of Measures of Centrality Based on Betweenness.” *Sociometry* 40 (1): 35–41. https://doi.org/10.2307/3033543.

Freeman, Linton C. 1978. “Centrality in Social Networks Conceptual Clarification.” *Social Networks* 1 (3): 215–39.

Gao, Song, Yaoli Wang, Yong Gao, and Yu Liu. 2013. “Understanding Urban Traffic-Flow Characteristics: A Rethinking of Betweenness Centrality.” *Environment and Planning B: Planning and Design* 40 (1): 135–53.

Garling, Tommy, Thomas Laitila, and Kerstin Westlin. 1998. *Theoretical Foundations of Travel Choice Modelling*. Pergamon.

Graser, Anita, Maximilian Leodolter, Hannes Koller, and Norbert Brändle. 2016. “Improving Vehicle Speed Estimates Using Street Network Centrality.” *International Journal of Cartography* 2 (1): 77–94.

Hong-cheng, Gan, Ye Xin, and Wang Qing. 2010. “Investigating the Effect of Travel Time Variability on Drivers’ Route Choice Decisions in Shanghai, China.” *Transportation Planning and Technology* 33 (8): 657–69.

Hoogendoorn, S. P., and P. H. L. Bovy. 2004. “Pedestrian Route-Choice and Activity Scheduling Theory and Models.” *Transportation Research Part B: Methodological* 38 (2): 169–90.

Jayasinghe, Amila, Kazushi Sano, Rattanaporn Kasemsri, and Hiroaki Nishiuchi. 2016. “Travelers’ Route Choice: Comparing Relative Importance of Metric, Topological and Geometric Distance.” *Procedia Engineering*, Proceeding of Sustainable Development of Civil, Urban and Transportation Engineering, 142 (January): 18–25.

Ji-hua, Hu, Huang Ze, and Deng Jun. 2013. “A Hierarchical Path Planning Method Using the Experience of Taxi Drivers.” *Procedia - Social and Behavioral Sciences*, Intelligent and Integrated Sustainable Multimodal Transportation Systems Proceedings from the 13th COTA International Conference of Transportation Professionals (CICTP2013), 96 (Supplement C): 1898–1909.

Kang, Chaogui, and Kun Qin. 2016. “Understanding Operation Behaviors of Taxicabs in Cities by Matrix Factorization.” *Computers, Environment and Urban Systems* 60 (November): 79–88.

Kaplan, Sigal, and Carlo Giacomo Prato. 2012. “Closing the Gap between Behavior and Models in Route Choice: The Role of Spatiotemporal Constraints and Latent Traits in Choice Set Formation.” *Transportation Research Part F: Traffic Psychology and Behaviour* 15 (1): 9–24.
Keay, Kevin, and Ian Simmonds. 2005. “The Association of Rainfall and Other Weather Variables with Road Traffic Volume in Melbourne, Australia.” *Accident Analysis & Prevention* 37 (1): 109–24.
Knight, Paul L., and Wesley E. Marshall. 2015. “The Metrics of Street Network Connectivity: Their Inconsistencies.” *Journal of Urbanism: International Research on Placemaking and Urban Sustainability* 8 (3): 241–59.
Knorring, J.H., R. He, and A.L. Kornhauser. 2005. “Analysis of Route Choice Decisions by Long-Haul Truck Drivers.” *Transportation Research Record*, no. 1923: 46–60.
Koller, H., P. Widhalm, M. Dragaschnig, and A. Graser. 2015. “Fast Hidden Markov Model Map-Matching for Sparse and Noisy Trajectories.” In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, 2557–61.
Liu, Chengxi, Yusak O. Susilo, and Anders Karlström. 2017. “Weather Variability and Travel Behaviour – What We Know and What We Do Not Know.” *Transport Reviews* 37 (6): 715–41.
Liu, Xiaoyue Cathy, Jeffrey Taylor, Richard J. Porter, and Ran Wei. 2018. “Using Trajectory Data to Explore Roadway Characterization for Bikeshare Network.” *Journal of Intelligent Transportation Systems* 0 (0): 1–17.
Lynch, Kevin. 1960. *The Image of the City*. MIT Press.
Maguire, Eleanor A., David G. Gadian, Ingrid S. Johnsrude, Catriona D. Good, John Ashburner, Richard S. J. Frackowiak, and Christopher D. Frith. 2000. “Navigation-Related Structural Change in the Hippocampi of Taxi Drivers.” *Proceedings of the National Academy of Sciences* 97 (8): 4398–4403.
Manley, E. J., J. D. Addison, and T. Cheng. 2015. “Shortest Path or Anchor-Based Route Choice: A Large-Scale Empirical Analysis of Minicab Routing in London.” *Journal of Transport Geography* 43 (February): 123–39.
Manley, E. J., S. W. Orr, and T. Cheng. 2015. “A Heuristic Model of Bounded Route Choice in Urban Areas.” *Transportation Research Part C: Emerging Technologies* 56 (Supplement C): 195–209.
Manley, E. J. 2016. “Estimating the Topological Structure of Driver Spatial Knowledge.” *Applied Spatial Analysis and Policy* 9 (2): 165–89.
Menghini, G., N. Carrasco, N. Schüssler, and K. W. Axhausen. 2010. “Route Choice of Cyclists in Zurich.” *Transportation Research Part A: Policy and Practice* 44 (9): 754–65.
Moar, I., and I. R. Carleton. 1982. “Memory for Routes.” *Quarterly Journal of Experimental Psychology*, no. 34A: 381–94.
Montello, Daniel R. 1998. “A New Framework for Understanding the Acquistion of Spatial Knowledge in Large-Scale Environments.” In *Spatial and Temporal Reasoning in Geographic Information Systems*, by M. J. Egenhofer and R. G. Golledge, 143–54. Oxford University Press.
Montello, Daniel R. 2005. “Navigation.” In *The Cambridge Handbook of Visuospatial Thinking*, by P. Shah and A. Miyake, 257–94. Cambridge University Press.
Oh, Ju Sam, Yong Un Shim, and Yoon Ho Cho. 2002. “Effect of Weather Conditions to Traffic Flow on Freeway.” *KSCE Journal of Civil Engineering* 6 (4): 413–20.
Pailhous, Jean. 1970. *La représentation de l’espace urbain: l’exemple du chauffeur de taxi*. Collection du travail humain. Paris: Presses Universitaires de France.
Papinski, Dominik, and Darren M. Scott. 2011. “A GIS-Based Toolkit for Route Choice Analysis.” *Journal of Transport Geography* 19 (3): 434–42.
Papinski, Dominik, and Darren M Scott. 2013. “Route Choice Efficiency: An Investigation of Home-To-Work Trips Using GPS Data.” *Environment and Planning A* 45 (2): 263–75.
Papinski, Dominik, Darren M. Scott, and Sean T. Doherty. 2009. “Exploring the Route Choice Decision-Making Process: A Comparison of Planned and Observed Routes Obtained Using Person-Based GPS.” *Transportation Research Part F: Traffic Psychology and Behaviour* 12 (4): 347–58.
Pas, Eric I., and Subramanian Sundar. 1995. “Intrapersonal Variability in Daily Urban Travel Behavior: Some Additional Evidence.” *Transportation* 22 (2): 135–50.

Porta, Sergio, Paolo Crucitti, and Vito Latora. 2006. “The Network Analysis of Urban Streets: A Primal Approach.” *Environment and Planning B: Planning and Design* 33 (5): 705–25.
Prato, Carlo Giacomo. 2009. “Route Choice Modeling: Past, Present and Future Research Directions.” *Journal of Choice Modelling* 2 (1): 65–100.
Prato, Carlo Giacomo, Shlomo Bekhor, and Cristina Pronello. 2012. “Latent Variables and Route Choice Behavior.” *Transportation* 39 (2): 299–319.
Ramaekers, Katrien, Sofie Reumers, Geert Wets, and Mario Cools. 2013. “Modelling Route Choice Decisions of Car Travellers Using Combined GPS and Diary Data.” *Networks and Spatial Economics* 13 (3): 351–72.
Regina O. Obe, and Leo S. Hsu. 2015. *PostGIS in Action*. Second edition. Shelter Island NY: Manning.
Stern, Eliahu, and Piet Bovy. 1990. *Route Choice: Wayfinding in Transport Networks*. Vol. Vol. 9. Studies in Operational Regional Science. Dordrecht: kluwer Academic Publishers.
Stern, Eliahu, and David Leiser. 1988. “Levels of Spatial Knowledge and Urban Travel Modeling.” *Geographical Analysis* 20 (2): 140–55.
Sun, Daniel (Jian), Chun Zhang, Lihui Zhang, Fangxi Chen, and Zhong-Ren Peng. 2014. “Urban Travel Behavior Analyses and Route Prediction Based on Floating Car Data.” *Transportation Letters* 6 (3): 118–25.
Sun, Longsheng, Mark H. Karwan, and Changhyun Kwon. 2016. “Incorporating Driver Behaviors in Network Design Problems: Challenges and Opportunities.” *Transport Reviews* 36 (4): 454–78.
Tang, Jinjun, Fang Liu, Yinhai Wang, and Hua Wang. 2015. “Uncovering Urban Human Mobility from Large Scale Taxi GPS Data.” *Physica A: Statistical Mechanics and Its Applications* 438 (November): 140–53.
Thomas, Tom, and Bas Tutert. 2015. “Route Choice Behavior in a Radial Structured Urban Network: Do People Choose the Orbital or the Route through the City Center?” *Journal of Transport Geography* 48 (October): 85–95.
Tomko, Martin, Stephan Winter, and Christophe Claramunt. 2008. “Experiential Hierarchies of Streets.” *Computers, Environment and Urban Systems* 32 (1): 41–52.
Venigalla Mohan, Zhou Xi, and Zhu Shanjiang. 2017. “Psychology of Route Choice in Familiar Networks: Minimizing Turns and Embracing Signals.” *Journal of Urban Planning and Development* 143 (2).
Wardrop, J. G. 1952. “Road Paper. Some Theoretical Aspects of Road Traffic Research.” *Proceedings of the Institution of Civil Engineers* 1 (3): 325–62.
Wiener, Jan M., Simon J. Büchner, and Christoph Hölscher. 2009. “Taxonomy of Human Wayfinding Tasks: A Knowledge-Based Approach.” *Spatial Cognition & Computation* 9 (2): 152–65.
Woollett, Katherine, and Eleanor A. Maguire. 2010. “The Effect of Navigational Expertise on Wayfinding in New Environments.” *Journal of Environmental Psychology* 30 (4): 565–73.
Yang, Lin, Mei-Po Kwan, Xiaofang Pan, Bo Wan, and Shunping Zhou. 2017. “Scalable Space-Time Trajectory Cube for Path-Finding: A Study Using Big Taxi Trajectory Data.” *Transportation Research Part B: Methodological* 101 (July): 1–27.
Yuan, J., Y. Zheng, X. Xie, and G. Sun. 2013. “T-Drive: Enhancing Driving Directions with Taxi Drivers’ Intelligence.” *IEEE Transactions on Knowledge and Data Engineering* 25 (1): 220–32.
Zhang, Shen, Jinjun Tang, Haixiao Wang, Yinhai Wang, and Shi An. 2017. “Revealing Intra-Urban Travel Patterns and Service Ranges from Taxi Trajectories.” *Journal of Transport Geography* 61 (May): 72–86.
Zhao, Shuangming, Pengxiang Zhao, and Yunfan Cui. 2017. “A Network Centrality Measure Framework for Analyzing Urban Traffic Flow: A Case Study of Wuhan, China.” *Physica A: Statistical Mechanics and Its Applications* 478 (July): 143–57.
Zhu, Shanjiang, and David Levinson. 2015. “Do People Use the Shortest Path? An Empirical Test of Wardrop’s First Principle.” *PLoS ONE* 10 (8).

Personal Declaration

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

June 30, 2018

Darryl Schumacher