University of Zurich[UZH]

Large-scale analysis of taxi drivers' route choice behaviour in San Francisco, Shanghai, and Vienna

GEO 511 Master's Thesis

Author
Tim Waldburger
12-918-322

Supervised by
Dr. Haosheng Huang
Anita Graser (anita.graser@ait.ac.at)

Faculty representative
Prof. Dr. Robert Weibel

28.06.2020
Department of Geography, University of Zurich

Abstract

Population growth and increasing individual mobility are putting the existing infrastructure in many urban areas under severe pressure. Successfully managing the expected traffic volumes of the next decades therefore requires progress in urban planning, traffic management, and alternative forms of mobility. An essential aspect in addressing these challenges is to understand individual route choice behaviour whereby the analysis of large datasets has played an increasingly important role in recent years. In this context, this thesis aims to investigate the route choice behaviour of taxi drivers using large-scale floating car data. The research is guided by the following three research questions: How do taxi drivers' routes differ from shortest, fastest, and fewest intersections routes? How does the street network impact taxi drivers' route choice behaviour? How do the answers to the above two research questions differ between San Francisco, Shanghai, and Vienna?

The analysis is based on three individual datasets from San Francisco, Shanghai, and Vienna and consists of three main steps: Firstly, the routes of over 2 million taxi trips are reconstructed from the floating car data. Secondly, for each of these routes, the shortest and fastest route as well as the route with the fewest intersections is computed. Thirdly, the taxi drivers' routes are compared to these optimal routes in terms of their similarity. Additionally, the routes are investigated in terms of road types, number of intersections and turns, as well as their spatial distribution within the street network.

The results reveal that most taxi drivers do neither follow the shortest, nor the fastest, or fewest intersections route but choose routes which are only partially congruent with these optimal routes. The extent to which the drivers follow one of these three optimal routes differs between the cities with drivers in Shanghai following them most often. Overall, taxi drivers' routes are about 10 % longer than the shortest route and 10–20 % slower than the fastest route. Furthermore, they include about 9 % more intersections than the fewest intersections route. On long trips, taxi drivers in Shanghai and Vienna tend to strongly prefer a short route over a route which is faster or includes less intersections. In contrast, drivers in San Francisco do not change their route choice behaviour even on long routes. The spatial distribution of taxi drivers' routes reveals that in all three cities,

the majority of trips is conducted within the city centre or between the airport and the centre. These are areas where the density and complexity of the street network is higher than in the other parts of the city, which means that taxi drivers have a large number of alternative routes to choose from and are therefore likely not to follow an optimal route.

These findings can help to improve traffic analysis and planning, resource management, map matching, and navigation systems. They may even provide insights into cultural differences in terms of route choice behaviour.

Keywords: route choice, floating car data, shortest path, route similarity, taxi drivers

Acknowledgements

Firstly, I would like to thank my main supervisor, Prof. Haosheng Huang, for guiding and supporting me during this exciting year of research. I would also like to thank Prof. Robert Weibel for offering me the opportunity of this project and Anita Graser for her valuable feedback during the writing of this thesis.

Additionally, I would like to thank Prof. Chun Liu from Tongji University and Taxi 31300 for providing the data which made this thesis possible in the first place. Furthermore, I want to point out that the street network data are copyrighted by OpenStreetMap Contributors and are available from https://www.openstreetmap.org as well as that the basemaps used in spatial visualisations are copyrighted by CARTO and by OpenStreetMap Contributors and are available from https://carto.com/.

Lastly, I would like to thank my partner for her patience and her unconditional support.

Contents

List of Figures

List of Tables

Acronyms and abbreviations

API Application Programming Interface. 28, 36, 40

BC Betweenness Centrality. 8, 74, 78, 101

CNL Cross-Nested Logit. 18

CPU Central Processing Unit. 36

DBMS Database Management System. 35

DTA Dynamic Traffic Assignment. 20

EBC Edge Betweenness Centrality. 7, 8, 37, 74–76, 78, 100, 101, 132

ESRI Environmental Systems Research Institute. 40

EWKB Extended Well-Known Binary. 134

FCD Floating Car Data. 1, 2, 11, 12, 14, 15, 21–24, 29, 33, 38, 75, 94, 100, 103, 107, 110, 111

FD Fréchet Distance. 25, 43, 63–67, 92, 93, 111, 132

FMM Fast Map Matching. 13, 40

GIS Geographic Information System. 36

GNL Generalized Nested Logit. 18

GNSS Global Navigation Satellite Systems. 1, 11–14, 39

GPS Global Positioning System. 22, 45, 46

HD Hausdorff Distance. 25

HMM Hidden Markov Model. 13

IDE Integrated Development Environment. 35

IQR Interquartile Range. 49, 50, 52, 56, 63, 65, 72, 88, 95

LBS Location-Based Services. 22

ML Machine Learning. 14, 20, 21

MM Map Matching. 1, 12, 13, 28, 36–40, 42, 45, 46

MNL Multinomial Logit. 17, 18, 23

MNP Multinomial Probit. 18

NL Nested Logit. 17–19

ODD Origin-Destination Distance. 55–57, 60–67, 70–74, 80–82, 84, 93–99, 102, 103

OSM OpenStreetMap. 28, 36–38, 41, 42, 78, 103, 111

PCL Paired Combinatorial Logit. 19

PID Percentage of Intersections Difference. 26, 43, 67, 68, 70–73, 98, 99, 132

PL Path-Size Logit. 19

PLD Percentage of Length Difference. 26, 43, 67, 68, 70, 72, 73, 98, 99, 132

PSL Percentage of Shared Length. 24–26, 43, 52–63, 65–67, 92–96, 111, 132

PTD Percentage of Time Difference. 26, 43, 67–73, 98, 99, 132

PVD Probe Vehicle Data. 11

RAM Random Access Memory. 36

RDS Relational Database Service. 36

RQ Research Question. 2, 3, 108, 109

SSSP Single-Source Shortest Path Problem. 10, 11

STA Static Traffic Assignment. 20

SVM Support Vector Machine. 20, 21

US United States. 30

Chapter 1

Introduction

1.1 Context and problem statement

At present, more than half of the global population lives in urban areas (United Nations 2018). In the future, these areas will absorb almost the entire population growth so that by 2030, they will be home to over 60 % of the world's inhabitants (United Nations 2019). In addition, over 2 billion people are expected to enter the middle class, leading to a doubling of today's global car fleet. As traffic conditions are already precarious in many cities, existing infrastructures will not be able to support this expected increase in traffic volume, resulting in unbearable costs due to time loss, health issues, and climate change effects caused by congestion (Dargay, Gately, and Sommer 2007; Bouton et al. 2015). In the last decade, research has therefore increasingly focused on improving resource management (Yuan et al. 2011, 2013), traffic analysis (Kim and Mahmassani 2015; D'Andrea and Marcelloni 2017; Zhang et al. 2019), and transportation networks (Ma et al. 2015, 2017). An essential aspect of these and other approaches addressing the challenge of future mobility is to understand individual route choice behaviour. To be able to identify and quantify the manifold factors influencing drivers' decisions and to predict which route they are most likely to choose is of key importance for the development of the future's transport systems (Sun et al. 2014; Lai et al. 2019) as route choice behaviour is an essential part in a variety of traffic-related applications such as traffic modelling, resource assignment, planning of infrastructure, and autonomous driving (Sun and Park 2017; Lai et al. 2019). Additionally, there are other applications which could benefit from a better understanding of individual route choice such as Map Matching (MM) and navigation systems (Miwa et al. 2012).

In the last two decades, the emergence of Floating Car Data (FCD) and further developments in data processing have enabled the empirical analysis of large Global Navigation Satellite Systems (GNSS) datasets which has become a widely used basis to analyse route

choice behaviour. Taxis are the preferred probe vehicles when generating FCD as they 1) are relatively cheap to equip with the necessary technology (Tang et al. 2015), 2) provide reliable trip information on a large scale (Wenk, Salas, and Pfoser 2006; Paulin and Bessler 2013), and 3) display more rational and heterogeneous route choice behaviour due to the above-average experience of the taxi drivers (Yao et al. 2013; Li, Wang, and Wang 2018).

A variety of studies based on taxi FCD has revealed that taxi drivers' route choice behaviour is influenced by many factors such as experience (Liu, Andris, and Ratti 2010), the presence of passengers (Nian, Zhu, and Sun 2017), as well as land use, and traffic composition (Peng et al. 2012). It has further been established that taxi drivers tend to optimise their routes but it remains unclear based on which criteria they optimise and to what extend they follow shortest paths (Li, Wang, and Wang 2018). Furthermore, the majority of related studies are based on data from Asian cities leaving other regions underrepresented. Additionally, there are only few studies comparing multiple cities to investigate differences in route choice behaviour (see Subsection 2.4.3).

1.2 Approach, research questions, and significance

Addressing the research gaps identified above, this thesis reveals, describes, and compares the movement patterns of taxi drivers in three major cities in North America, Asia, and Europe. It is based on the assumption that the drivers' route choice behaviour is reflected by the characteristics of the routes they choose (Taylor, Gardony, and Brunyé 2018) and therefore empirically analyses three large sets of FCD collected from taxis in San Francisco, Shanghai, and Vienna. The analysis is divided into three main parts, namely 1) the extraction of the routes actually chosen by taxi drivers from the data, 2) the computation of the shortest, fastest, and fewest intersections[1] alternatives for each actual route, and 3) the comparison of the results between the three cities and the different route types. The thesis is guided by the following Research Questions (RQs):

RQ1 How do taxi drivers' routes differ from shortest, fastest, and fewest intersections routes?

- 1.1 Do taxi drivers with passengers on board take the shortest, fastest, or fewest intersections routes?
- 1.2 How much longer is the actual route than the shortest route, how much slower is the actual route than the fastest route, and how many more intersections does the actual route include than the fewest intersections route?

[1] Throughout this thesis, the term "fewest intersections route" refers to the route which includes the fewest intersections when travelling from a given origin to a given destination.

RQ2 How does the street network impact taxi drivers' route choice behaviour?

- 2.1 Is there significant correlation between edge betweenness centrality and the routes chosen by taxi drivers with passengers on board?
- 2.2 Do taxi drivers with passengers on board avoid or prefer particular road types?
- 2.3 Do taxi drivers with passengers on board avoid or prefer complex intersections?
- 2.4 Do taxi drivers with passengers on board avoid or prefer right or left turns?

RQ3 How do the findings from RQs 1 and 2 differ among San Francisco, Shanghai, and Vienna?

The findings presented in this thesis will contribute towards a more comprehensive understanding of taxi drivers' route choice. The identification of influencing route characteristics and the quantification of the extent to which taxi drivers follow optimal routes[2] will provide a basis for further developments in route choice analysis and modelling. The findings on how the street network affects route choice behaviour will be useful for the planning of traffic networks and infrastructure as well as to apply findings gained in one city to other cities. Furthermore, the identified research gap regarding a lack of studies comparing multiple cities as well as the lack of related research outside Asia is addressed by the combination and comparison of multiple datasets from three different cities on three different continents. Assessing similarities and differences between these cities will also provide a basis to transfer knowledge gained by previous research. As driving practice varies in different cultures (Özkan et al. 2006), the comparison of data collected by taxis in three different cultural areas might also present new insights into cultural differences in route choice behaviour.

1.3 Structure of the thesis

The remainder of this thesis is structured as follows: a theoretical and technical background is provided in Chapter 2 whereby the current research and the research gaps are outlined in subsections 2.4.3 and 2.4.5. Chapter 3 then introduces the data and the study areas. The methodological approach is outlined in Chapter 4. Chapter 5 presents the results of the analysis which are then discussed in Chapter 6. A synthesis answering all RQs is presented in Section 6.3. Finally, the thesis is concluded in Chapter 7.

[2] In this thesis, the term "optimal routes" refers to the shortest, fastest, and fewest intersections routes. These routes show minimal cost according to a predefined criterion, namely length, duration, and number of intersections.

Chapter 2

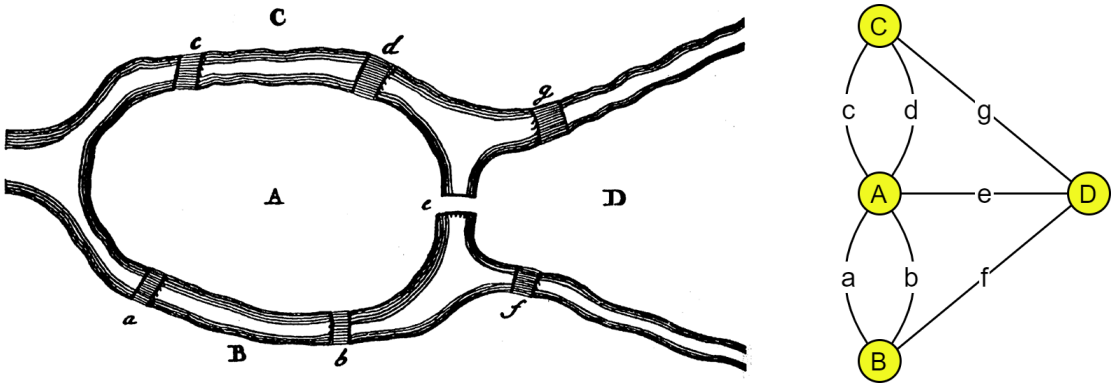
Theoretical and technical background

2.1 Graph theory and network science

2.1.1 Historical background

In geographic information science, and especially in disciplines dealing with networks, concepts from graph theory find many useful applications (George 2016). Graph theory as a branch of mathematics was founded by Leonhard Euler who published an article on a known mathematical problem, called the "Königsberg bridge problem", in 1741 (Euler 1741). The problem referred to the city of Königsberg in former Prussia[3], which was divided into two islands and two banks by the Pregel River with seven bridges connecting the different parts of the city (see Figure 2.1a). The problem was formulated based on this specific city layout and asked for a round trip that starts at any point and crosses each bridge exactly once. Euler ended the dispute of then mathematicians by proving that such a path did not exist (George 2016). He came to his conclusion by abstracting the situation into a graph - a system of points (called nodes or vertices) and lines (called edges or links), which strongly abstracts the original problem but still represents all its connectivity (Newman, Barabási, and Watts 2006a; Diestel 2017) (see Figure 2.1b). Euler argued that the path required by the problem occurs only in graphs in which all nodes have an even degree (see Subsubsection 2.1.3.1) and that it can therefore not be found for Königsberg since its corresponding graph has four nodes of odd degree (see Figure 2.1b). This proof is widely considered as the beginning of graph theory which has become the essential mathematical tool to describe networks and their properties (Barnes and Harary 1983; West 1996). In our modern world, networks and problems that can be described as such

[3] Today, Königsberg is known as Kaliningrad and part of Russia.



(a) Diagram from Euler's 1741 paper (b) Graph

Figure 2.1: The bridges of Königsberg represented as a diagram (a) and as the corresponding graph (b). The landmasses are named with capital letters and the bridges are labelled in lowercase. Figure modified from Hopkins and Wilson (2004, p.199–200).

are omnipresent, which is why in the last decades mathematicians, biologists, sociologists and others have been contributing towards the emergence of the new research field of network science (Barabási 2002; Buchanan 2002; Newman, Barabási, and Watts 2006b). According to the National Research Council (2005, p.28), network science "consists of the study of network representations of physical, biological, and social phenomena leading to predictive models of these phenomena". In network science, a network and an (attributed) graph are used as synonyms (Barthélemy 2014)[4]. A graph can represent almost anything: spread of disease, human organisations, social interrelations, and transportation as well as flows of information, capital, goods, and individuals (National Research Council 2005). A graph is usually defined as $G = (N, E)$, where N is a set of nodes and E is a set of edges connecting the nodes. In weighted graphs, l describes the edge weights (Barthélemy 2018).

2.1.2 Spatial graphs

Spatial graphs are graphs in which the nodes are located in a metric space whereby most applications use a 2–dimensional Euclidean space (Barthélemy 2014). While non-spatial graphs are usually sufficiently characterised by an adjacency matrix defining the graph's topology[5], spatial graphs require additional spatial information, such as coordinates, for

[4] In this thesis, "graph" is used as a technical term to describe structured and attributed data while "network" refers less specifically to a general structure of connected nodes and edges.

[5] For a given graph with n nodes, the corresponding $n \times n$ adjacency matrix stores information about which nodes are connected by an edge. The cell A_{ij} contains the value 1 if the nodes i and j are connected and the value 0 if there is no edge connecting them. In weighted networks, the weight value is stored instead (Barthélemy 2018).

a full description, which is usually encrypted in the nodes' locations. Node and edge attributes in spatial graphs are dependent on spatial aspects, such as edge lengths, implying that the probability of a connection between two nodes changes with the distance separating them (Barthélemy 2011, 2014). The possibility of two topologically identical graphs having different spatial properties (see Figure 2.2) is the reason for both – their versatility and their complexity. For many networks representing real world phenomena, space is relevant because topology alone does not contain all information; people tend to have more social contacts in their neighbourhood and travel times in transportation networks depend on distance. The importance to include space in networks has already been discussed several decades ago (Haggett and Chorley 1969) and models to characterise spatial networks have been developed (Chorley and Haggett 1967). In geography, advances in spatial networks analysis contributed towards a better understanding of the spatial structure of urban areas, human mobility, and similar (Barthélemy 2011). Like many other scientific disciplines, network science has and will greatly benefit from an increasing availability of network data and computational power (Vespignani 2018).

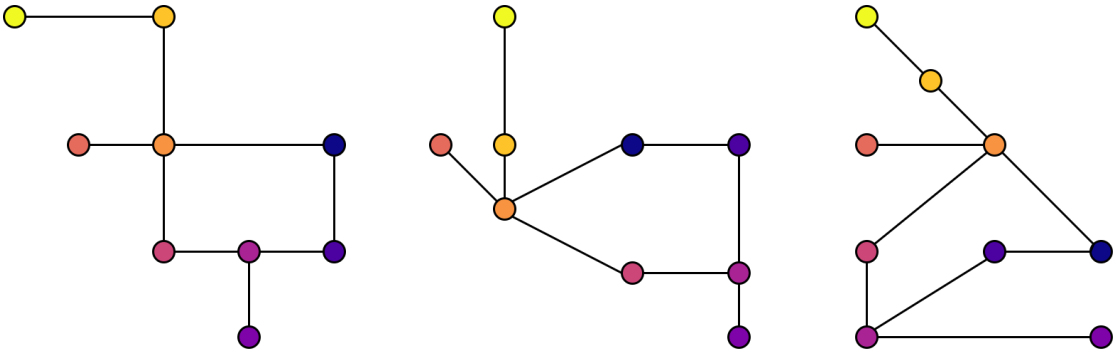


Figure 2.2: The networks represented by these three graphs all have the same adjacency matrix and are topologically equivalent but the spatial representations vary. The differing spatial information is encoded in the nodes' positions. Figure modified from Barthélemy (2018, p.4).

Street networks represent a special form of spatial networks because in general, but especially in urban areas, they are characterised by strong planarity[6] and a defining influence of metric distances. Nevertheless, they share many similarities with non-spatial networks (Crucitti, Latora, and Porta 2006; Porta, Crucitti, and Latora 2006a). In graphs representing street networks, streets are usually represented by edges and intersections between them are represented as nodes. Both can then be attributed with a variety of

[6] In planar graphs, edges are only allowed to intersect at nodes (Rodrigue, Comtois, and Slack 2017). In street networks, planarity might be violated because of underpasses, tunnels, and bridges but planar graphs are still considered a good approximation (Lämmer, Gehlsen, and Helbing 2006).

characteristics, such as speed limit, road pavement, or the presence of light signalling devices. Street networks are of use when investigating the influence of spatial layouts on social, environmental, and economic phenomena (Hillier and Hanson 1984) because their analysis allows to capture spatial characteristics of cities as well as human movement patterns (Barthélemy 2011) and therefore provides new opportunities for urban designers and traffic planners (Porta, Crucitti, and Latora 2006b).

2.1.3 Network measures and indices

Comparing complex networks is challenging and therefore, various measures and indices have been developed to quantify a network's efficiency and accessibility. For street networks, they specifically aim at comparing different networks at a specific point in time or at analysing a single network's evolution over time (Rodrigue, Comtois, and Slack 2017). Since a detailed summary would go beyond the scope of this work, the remainder of this section focuses on two measures that are central to this thesis: 1) node degree and 2) Edge Betweenness Centrality (EBC). Nevertheless, a rough overview of network measures is presented in Table 2.1 at the end of the section.

2.1.3.1 Node degree

Node degree (or order) describes the importance of a node in the network. It is defined by the number of edges attached to the node (see Figure 2.3). A high value indicates that the node is important since it means that many edges are attached to it. In directed graphs[7], in- and out-degree are distinguished as the number of edges going in or out of the node and differences might indicate a node's function as attractor or sender (Rodrigue, Comtois, and Slack 2017).

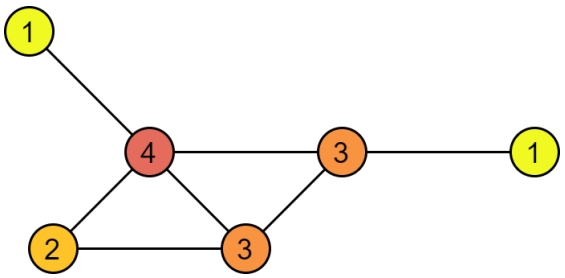


Figure 2.3: Graph with node degree assigned to each node.

[7] In directed graphs, edges can only be traversed in one direction whereby in undirected graphs, they can be traversed in both directions (Diestel 2017).

2.1.3.2 Edge betweenness centrality

Besides a multitude of centrality indicators, Betweenness Centrality (BC) is a widely used accessibility measure to describe the importance of a node or edge and to capture the structure of a network. In contrast to complex non-spatial networks where BC scales with the degree, BC and degree are usually not equivalent in spatial networks which makes the distribution pattern of BC a valuable resource of information about a network's relations between topology and space (Rodrigue, Comtois, and Slack 2017; Barthélemy 2018). BC can be calculated for nodes and edges but as this thesis only uses EBC, this section will focus on the latter.

As Equation 2.1 shows, EBC is based on the number of shortest paths travelling through an edge. It can be defined as:

$$g(e) = \frac{1}{\mathcal{N}} \sum_{s \neq t} \frac{\sigma_{st}(e)}{\sigma_{st}}, \tag{2.1}$$

where σ_{st} is the number of shortest paths from s to t and $\sigma_{st}(e)$ is the number of shortest paths from s to t travelling through edge e. $\mathcal{N}$ is the normalisation constant to ensure that $g(e) \in [0, 1]$. It represents the number of node pairs in the graph and in directed graphs, it is calculated using $\mathcal{N} = (N-1)(N-2)$, where N is the number of nodes in the graph (Barthélemy 2018). In summary: the more shortest paths pass through an edge, the greater its importance and the larger its EBC. Since the numerical calculation of EBC requires the computation of shortest paths between each pair of nodes in a graph, it can be computationally intensive for large and complex networks. The standard algorithm to calculate BC was presented by Brandes (2001) and runs in $\mathcal{O}(NE)$ time for unweighted and in $\mathcal{O}(NE + N^2 \log N)$ time for weighted graphs, where N is the number of nodes and E is the number of edges in the graph (Brandes 2001; Barthélemy 2018).

Table 2.1: Network measures and indices. The first column names the characteristic described by the measures in the second column. The third column lists related literature.

At network level		
Extent	Diameter, length of segments, eta index	1,4
Structure	Shape, cell area, transitivity, hierarchy, assortative coefficient, organic ratio, degree distribution	1,2,3,4,5
Complexity	Number of cycles, density, pi index, theta index	1
Connectivity	Alpha index, beta index, gamma index	1,2,3
Efficiency	Cost, performance, efficiency, detour index, average shortest path length	1,2,3,4,6
At node/edge level		
Importance	Degree, betweenness centrality, straightness centrality, information centrality, degree centrality	1,2,3,4,5,6,7
Vulnerability	Hub dependence, participation coefficient	1
Independence	Closeness centrality	6,7
Accessibility	Local/global efficiency, Shimbel index, Koenig number	1,5,6
Clusters	Clustering coefficient, average nearest neighbour degree, cohesion index, z-score, assortativity	1,3,4,5

1 Rodrigue, Comtois, and Slack (2017)
2 Barthélemy (2011)
3 Barthélemy (2014)
4 Barthélemy (2018)
5 Porta, Crucitti, and Latora (2006a)
6 Porta, Crucitti, and Latora (2006b)
7 Crucitti, Latora, and Porta (2006)

2.1.4 Shortest path search

Shortest path problems are fundamental in graph theory and of great importance in transportation science. Transportation problems regularly require finding shortest paths between origin and destination nodes in a graph whereby the cost to be minimised may differ (Pallottino and Scutellà 1998).[8] Since determining a shortest path requires finding

[8] Criteria to determine shortest paths might be distance, travel time, number of signposts, or cost (Bovy and Stern 1990).

the shortest paths from a given origin node to all other nodes in the graph, this problem can be formalised as finding the shortest paths from a given origin node $o \in N$ to every node $n \in N$ in a given graph $G = (N, E, l)$, where N and E represent the nodes and edges in G and l represents the length function by which the shortest path is determined. This problem is called Single-Source Shortest Path Problem (SSSP) (or simply shortest path problem) and its complexity primarily depends on whether edge weights are positive or negative and whether the graph is directed or undirected (Pettie 2008). The remainder of this subsection briefly introduces some fundamental algorithms proposed to find shortest paths under different conditions. For a comprehensive survey on shortest-path computation see Pallottino and Scutellà (1998).

One of the most popular algorithms in computer science (Sniedovich 2006), and arguably the most famous shortest path algorithm, is Dijkstra's algorithm (Dijkstra 1959). It solves the SSSP for directed and undirected graphs with positive edge weights. The algorithm starts from the specified origin node and visits all other nodes. Nodes which have not been visited yet are kept in a priority queue and the fundamental idea is to always process the closest node next. Dijkstra's algorithm is a so-called greedy algorithm, which means that it always prefers the currently most promising solution and continues without backtracking (Sammut 2011). This method ensures that the first path reaching a node is always the shortest path. The algorithm is finished once all reachable nodes have been visited achieving a time complexity of $\mathcal{O}(N^2)$, where N is the number of nodes in the graph (Madkour et al. 2017; Bhatia 2019).

Similar to Dijkstra's algorithm, the Bellmann-Ford algorithm (Ford 1956; Bellman 1958; Moore 1959) also solves the SSSP by computing the shortest paths between an origin node and all other nodes in the graph. Both algorithms work similarly but instead of selecting the closest node, the Bellmann-Ford algorithm selects all adjacent edges. Although it is slower than Dijkstra's algorithm, it has the advantage of being able to handle negative edge weights. Negative weights mean that cycles with negative weight in the graph must be found because otherwise they could be traversed multiple times constructing a new shortest path in each iteration. The Bellmann-Ford algorithm is able to identify and deal with negative cycles and runs in $\mathcal{O}(NE)$ time, where N is the number of nodes and E is the number of edges in the graph (Wong and Tam 2005; Bhatia 2019).

The Floyd-Warshall algorithm (Floyd 1962; Warshall 1962) is different as it does not calculate shortest paths from a single origin but between all node pairs in a directed graph. It follows the principle of dynamic programming (Rust 2016) and is based on the idea that a shortest path between the nodes o and d going through node n also includes the shortest paths from o to n and from n to d. Although the algorithm can detect negative cycles, it is not able to resolve them, meaning that graphs with negative edge weights

can be processed only if no negative cycles are present. The Floyd-Warshall algorithm's time complexity is $\mathcal{O}(N^3)$, where N is the number of nodes (Madkour et al. 2017).
For sparse graphs[9], this time can be improved using Johnson's algorithm (Johnson 1977) which transforms the graph by removing negative edge weights using the Bellmann-Ford algorithm and then runs Dijkstra's algorithm on the transformed graph.
When working with FCD, the underlying street network can usually be represented as a weighted directed graph. If the graph's edge weights are positive, it fulfils all requirements for Dijkstra's algorithm and if edge weights contain negative values, the Bellmann-Ford algorithm can be used in order to solve the SSSP (Peixoto 2020).

2.2 Floating car data

Over the last two decades, Floating Car Data (FCD), synonymously called Probe Vehicle Data (PVD)[10], has become a powerful tool in the assessment of urban traffic conditions. The term refers to data acquired by sample vehicles in order to represent overall traffic conditions and was popularised by Schäfer, Thiessenhusen, and Wagner (2002). The principle is illustrated by the analogy of a cork swimming in the river where one can derive the river's flow rate and direction by measuring the cork's direction and velocity (Pfoser 2008; Pfoser et al. 2008). FCD is generated by taxis, public transport vehicles and utility vehicles, as well as by private vehicle owners (Messelodi et al. 2009). A GNSS-enabled device periodically sends its location and an exact time stamp whereby the data points are usually attributed with additional information, such as driving speed or direction of travel, and sent to a central processing server. After pre-processing, the data are stored in a database as spatial trajectories (Pfoser 2008; Pfoser et al. 2008; Paulin and Bessler 2013; Zheng 2015). FCD proves itself superior to survey-based techniques because it is cheaper and less time consuming. It provides reliable trip information for large regions and allows to observe the actual routes chosen by drivers on a large scale (Wenk, Salas, and Pfoser 2006; Paulin and Bessler 2013). In the case of route choice behaviour, public transport and utility vehicles are unsuited as probe vehicles because they drive on predefined routes without making their own route choices. Taxis as sensors to collect FCD possess some particular advantages over private vehicles: there is no additional cost for hard- and software because taxis are usually already equipped with GNSS receivers and wireless communication devices, their number is large enough to cover an entire city's street network, and they are better suited than mobile phone data

[9] Sparse graphs are graphs where the relation between the number of nodes and the number of edges is about linear (Diestel 2017).
[10] Although the terms "FCD" and "PVD" are used as synonyms (Pfoser et al. 2008), "FCD" is preferred in this thesis.

to represent passengers' origins and destinations (Kühne et al. 2003; Tang et al. 2015). Taxi drivers are also considered to possess more driving experience and better knowledge about road traffic conditions than ordinary drivers. This results in more rational and more heterogeneous route choice behaviour (Yao et al. 2013; Li, Wang, and Wang 2018).

2.3 Map matching

There are two main sources of error when collecting FCD: GNSS-related positioning errors and position uncertainties at low sampling rates, because the position of the vehicle between data points is ambiguous (Pfoser and Jensen 1999). Both errors can be corrected by matching the tracking data to the underlying street network. This process of identifying the correct location of a vehicle on a network edge is called Map Matching (MM) (Quddus, Ochieng, and Noland 2007; Jensen and Tradišauskas 2009). There are two main approaches which fundamentally differ in terms of the error source they address: online MM and offline MM. Most algorithms perform online MM, meaning that individual points need to be matched onto line segments in real-time which then again implies that only past positions are available. Online MM is heavily used in on-board navigation systems aiming at minimizing GNSS-related positioning errors. Offline MM algorithms on the other hand, are mostly concerned with uncertainties caused by low sampling rates. When performing offline MM, not single points but entire trajectories are matched. Therefore, not only past but all positions are available to match a given point allowing the algorithm to process more data and to produce better results (Brakatsoulas et al. 2005; Jensen and Tradišauskas 2009). Offline MM is used when the collected data do not need to be processed in real-time.

The algorithms used when map matching FCD can be classified into geometric, topological, probabilistic, and advanced algorithms (Quddus, Ochieng, and Noland 2007; Chen, Shen, and Tang 2011). Geometric algorithms only use the geometric information contained in the street network, such as the shape of the roads. Because they use only limited data, they are fast and simple but also error-prone when dealing with intersections, roundabouts, and parallel streets (Velaga, Quddus, and Bristow 2009). Topological methods consider geometry, connectivity, and adjacency of the street network segments and are therefore more accurate than geometric approaches. However, they do not use information about vehicle speed and heading and are therefore sensitive to outliers (Zhihua and Wu 2005; Chen, Shen, and Tang 2011). Probabilistic algorithms use network topology, past and future vehicle locations, information about the positional error, and additional vehicle data to define an error region around a given GNSS data point. They then select multiple street segments in this error region as possible solutions and choose

the most likely candidate (Ochieng, Quddus, and Noland 2003). Advanced MM methods make use of Kalman filtering, fuzzy logic, evidence theory, and neural networks. They are superior to the other approaches but also more complicated, slower and require more input data (Quddus, Noland, and Ochieng 2006; Chen, Shen, and Tang 2011). The selection of a suitable MM algorithm is based on the size of input data, the desired quality of the results, as well as the computational capacities available. The algorithm used in this thesis is a stochastic offline algorithm and uses advanced methods such as Hidden Markov Models (HMMs) (see Subsection 2.3.1).

Many MM algorithms from the last decade use HMMs (Rabiner and Juang 1986) to model network topology, GNSS error, and other parameters. HMMs have been used in speech processing and bioinformatics for decades but were only introduced to map matching in 2009 (Newson and Krumm 2009; Lou et al. 2009). Substantial effort has been made to improve MM algorithms by processing more data (Zheng et al. 2016), incorporating shortest path searches (Wang et al. 2011), and implementing concepts of drivers' route choice behaviour (Miwa et al. 2012; Hunter, Abbeel, and Bayen 2014). The problem of increasingly large datasets requiring high-performance MM algorithms has been addressed by Koller et al. (2015) as well as by Yang and Gidófalvi (2018).

2.3.1 Fast map matching algorithm

Fast Map Matching (FMM) is an algorithm specifically designed to achieve high performance and was proposed by Yang and Gidófalvi (2018). Before the actual map matching, an origin-destination table storing all shortest paths under a given length is created from the street network using Dijkstra's algorithm (Dijkstra 1959) (see Subsection 2.1.4). During MM, repeated and computationally intensive shortest path searches are then replaced with faster hash table searches in the precomputed paths. Yang and Gidófalvi (2018) optimised and parallelized their implementation in C++ to run as fast as possible.[11] The authors claim to achieve performances several times higher than other algorithms (Yang and Gidófalvi 2018).[12] Similar to other MM approaches, FMM uses a HMM taking into account topological constraints and GNSS error. It also introduces a penalty for reverse movement since a vehicle repeatedly travelling back and forth on two directed edges is most likely a result of GNSS error and not an actual driving behaviour (Yang and Gidófalvi 2018). FMM takes a set of trajectories and a network as input and returns the travelled edge ids and the geometry for each trajectory together with a whole set of additional attributes. In terms of matching percentages, the results greatly

[11] The algorithm's open-source implementation is available at https://github.com/cyang-kth/fmm.
[12] An average matching speed of 35'000 points per second was achieved during the MM for the present work.

depend on the set of input parameters selected by the user and on the accuracy of the GNSS data but match percentages of over 90 % are achievable (Yang and Gidófalvi 2018).

2.4 Route choice behaviour

2.4.1 Historical background

After Wardrop (1952) had introduced the user equilibrium (see Subsubsection 2.4.2.1), shortest paths (see Subsection 2.1.4) were believed to be the foundation of route choice for a long time. It was also assumed that all traffic participants know the entire street network as well as the current traffic conditions and that they therefore rationally choose the best route whereby the lowest perceived cost was usually taken as criterion for route selection (Willumsen 2000; Morikawa et al. 2005). However, research has shown that shortest paths alone show poor performance in predicting route choice due to drivers' imperfect network knowledge, cognitive limitations, intrinsic preferences, and bounded rationality (Liu and Xu 2019). An alternative was presented by Luce (1959) who proposed a stochastic modelling approach which incorporated uncertainty in route choice using probability functions providing a basis for a range of further developments and adjustments (see Subsubsection 2.4.2.2). While continuous progress has been made in modelling, only the increasing availability of data and computing capacity, as well as the development of Machine Learning (ML) methods, have recently opened up new possibilities in transportation analysis (Miyagi 2004; Kruppa et al. 2013; Gupta and Pathak 2014; Jahangiri and Rakha 2015; Tang et al. 2017). Although relatively little effort has been made to use ML for the analysis of route choice behaviour (Lai et al. 2019), advanced ML techniques have shown promising results this particular field of research (Henn 2003; Wei, Ma, and Jia 2014; Sun and Park 2017; Lai et al. 2019).

According to COMSIS Corporation (1995), influencing factors of route choice include level of experience, current traffic conditions, trip purpose as well as perception of alternative routes, time or monetary cost, and comfort. Lack of data was one of the reasons why not all these factors could be verified (Morikawa et al. 2005). Apart from origin-destination matrices and traffic counts, sources of information about route choice behaviour were surveys, field experiments, route choice modelling, and interactive computer simulation games whereby field experiments presented the most accurate results. However, field experiments were often limited to small sample sizes because they were time-consuming and expensive (Morikawa et al. 2005). After the turn of the millennium, the emergence of FCD opened up a new source of data (see Section 2.2) providing reliable trip information on a large scale (Wenk, Salas, and Pfoser 2006; Paulin and Bessler

2013). FCD, in combination with increasing computation capacities, has produced new findings in route choice analysis, namely about the relation between street network and route choice (Jiang, Yin, and Zhao 2009; Liu et al. 2015), traffic flow composition and trip purpose (Peng et al. 2012), as well as the factors influencing route choice. The assumption about drivers always choosing the shortest route has been rejected and replaced with new hypotheses (Liu, Andris, and Ratti 2010; Yao et al. 2013; Manley, Addison, and Cheng 2015; Li, Wang, and Wang 2018) whereby some of the influencing factors presented by COMSIS Corporation (1995), namely level of experience and network knowledge, have been verified (Liu, Andris, and Ratti 2010).

2.4.2 Route choice modelling

Following Wardrop (1952), a variety of different route choice modelling approaches has been proposed addressing drivers' incomplete network knowledge, personal preferences, bounded rationality, and more. This section does not claim completeness but provides an overview of different approaches in route choice modelling in a loosely chronological order. Comprehensive reviews including mathematical formulations of the models are provided by Ben-Akiva, Ramming, and Bekhor (2004), Prashker and Bekhor (2004), and Prato (2009).

2.4.2.1 Deterministic models

Deterministic route choice models do not include any randomness and the results therefore depend entirely on the model's parameters. They are relatively trivial to formulate and implement but also only able to reflect reality to a very limited extent (Rey 2015). Deterministic models have a long history in route choice modelling. In the first half of the 20$^{\text{th}}$ century, traffic planners' only tool to model traffic and congestion were empirical relationships (Greenshields et al. 1935). The missing piece was eventually provided by Wardrop (1952) who formulated the math, providing the basis to model route choice behaviour, with the following two principles at its core:

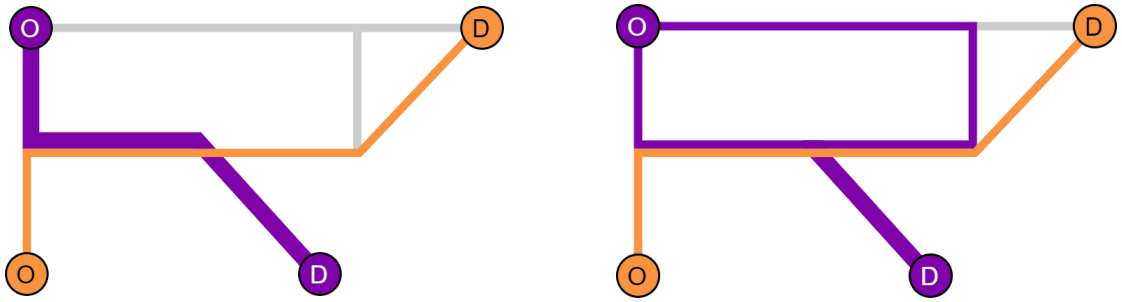
(I) "The journey times on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route (Wardrop 1952, p.345)"

(II) "The average journey time is a minimum (Wardrop 1952, p.345)."

Under the first principle, referred to as user equilibrium[13], every user unilaterally min-

[13] Wardrop's user equilibrium is related to the Nash equilibrium (Nash 1950, 1951; Osborne and Rubinstein 1994) which is one of the fundamental concepts in game theory. Although they apply different solution concepts, the Wardrop's user equilibrium can be formulated as an instance of a Nash equilibrium (Correa and Stier-Moses 2011).

imises their individual travel time without taking into account the other traffic participants (see Figure 2.4a). Under user equilibrium, no user is able to reduce their travel time and the equilibrium therefore persists as long as travel demand in the network is unchanged (Chiu et al. 2011). In contrast, under Wardrop's second principle, called the system optimal, each user aims at minimising the total travel time in the network. This means, that traffic participants avoid congestion even if it leads to longer travel times for themselves (see Figure 2.4b). Wardrop's user equilibrium became a fundamental behavioural assumption in early route choice modelling and transportation planners have been, and still are, using it for real-life applications (Sheffi 1985; Florian 1999; Correa and Stier-Moses 2011).



(a) User equilibrium (b) System optimal

Figure 2.4: The figure represents traffic flows from origin (O) to destination (D). Under user equilibrium (a), each traffic participant chooses the shortest route resulting in congestion on this road. Under system optimal conditions (b), some participants take a detour to avoid congestion resulting in evenly distributed traffic without congestion. Figure modified from Wen (2019, p.1).

In deterministic traffic modelling, finding the shortest path between two given nodes is the simplest form of a route choice model (Ben-Akiva, Ramming, and Bekhor 2004). Shortest path search is so fundamental that most existing research on path computation has been dedicated to this problem (Zhang 2017). Initial deterministic approaches were proposed by Bellman (1958) and Dijkstra (1959) whose work was then extended in heuristic (Hart, Nilsson, and Raphael 1968; Kung et al. 1984) and hierarchical (Jing, Huang, and Rundensteiner 1998) ways. Since shortest path search is a central part of this thesis, the topic is introduced in more detail in Subsection 2.1.4.
Another deterministic method is the so-called labelling approach (Frejinger, Bierlaire, and Ben-Akiva 2009). It assumes that different drivers try to optimise their routes based on different criteria, such as shortest, fastest, or cheapest route, and that each criterion might lead to a different preferred route. Based on this assumption, each route can now be labelled according to the criteria for which it is the optimal route (Bovy and Stern

1990; Prato 2009). The labelling approach is used in combination with other models such as shortest path or Nested Logit (NL) (Ben-Akiva et al. 1984; Ramming 2002).
Link elimination and link penalty models are further approaches in deterministic route choice modelling: link elimination repeats a shortest path search and constantly removes the shortest path links from previous searches. This guarantees that alternatives do not unintentionally share links. The rules according to which links are eliminated can be specified by the user. Link elimination has some flaws because it might lead to disconnected networks which causes the shortest path algorithm to miss potentially optimal routes (Prato 2009). Equivalent to link elimination, the link penalty approach (De La Barra, Pérez, and Añez 1993; Ruphail et al. 1996) is also based on repeating shortest path searches but instead of removing links, a penalty on the impedance of all shortest path links is imposed. This has the advantages that more dissimilar routes are generated and that potentially essential links remain in the network. However, the generation of paths with high impedance might negatively affect the relevancy of the generated routes (Prato 2009).

2.4.2.2 Stochastic models

The deterministic approach is based on the assumption that travellers possess perfect knowledge of the street network and the current traffic conditions as well as on an identically perfect rational behaviour of all traffic participants. However, these assumptions have been criticised for not reflecting real-life conditions (Gärling 1998; Manley, Orr, and Cheng 2015). The stochastic approach therefore tries to model drivers' incomplete network knowledge using probability distributions (Daganzo and Sheffi 1977). These distributions represent the probability of a driver taking a particular link when moving from one node to another.
The Multinomial Logit (MNL) model (Luce 1959) represents an alternative to deterministic models and uses the Gumbel distribution to model traffic participants' incomplete network knowledge. Due to its simple mathematical formulation, it has become the most widely used choice model (Wen and Koppelman 2001). However, the MNL model is not suitable for route choice modelling because it does not account for similarity among alternative routes (Prashker and Bekhor 2004). The problem is that there are up to 100 alternative routes with high similarity in dense urban networks and that the MNL model is not able to accurately represent the probability that a certain route is chosen when there are overlapping routes (see Figure 2.5) (Prato 2009).

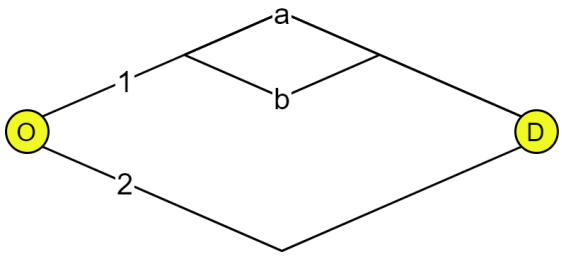


Figure 2.5: Assume that the three routes (1a, 1b, 2) from origin (O) to destination (D) have equal length and that distance is the only attribute considered. In this case, the MNL model provides a probability of 0.33 for each route although, given that 1a and 1b are so similar, we would expect probabilities of around 0.25 for 1a and 1b and a probability of around 0.5 for route 2. Figure modified from (Ben-Akiva and Bierlaire 2003, p.13).

The NL model (Ben-Akiva 1973, 1974) is an extension of the MNL model designed to overcome its drawbacks by dealing with route similarity. It is tree-structured and partitions route alternatives into nests whereby alternatives in each nest are correlated. However, the NL model cannot capture correlation across nests and its tree-structure also assumes that each route exclusively belongs to one nest while in real-life networks, routes share edges with many other paths. The NL model is therefore not suitable to model route choice (Ben-Akiva and Bierlaire 2003; Prashker and Bekhor 2004; Prato 2009).

The Cross-Nested Logit (CNL) (or Generalized Nested Logit (GNL)) model (McFadden 1978; Vovsha and Bekhor 1998; Wen and Koppelman 2001) extends the NL model in a way that allows alternative routes to belong to more than one nest (Ramming 2002). The probability of choosing a certain route depends on its utility, an inclusion coefficient which is similar to the probability function in the MNL model, and the nesting coefficient. The CNL model therefore collapses into the MNL model if the nesting coefficient is equal to one (Ben-Akiva, Ramming, and Bekhor 2004).

The Multinomial Probit (MNP) model (Daganzo and Sheffi 1977) explicitly takes into account the correlation between all route alternatives and models the random component in human route choice behaviour using a normal distribution. Because of this, the computation of the MNP model becomes very demanding when the number of alternative routes is larger than three (Bovy and Stern 1990; Ramming 2002; Prashker and Bekhor 2004). Despite efforts to reduce processing costs (Sheffi and Powell 1982; Yai, Iwakura, and Morichi 1997), the MNP model's high computational demands are the reason why alternative modelling approaches are often preferred in route choice modelling (Prato 2009).

The C-Logit model (Cascetta et al. 1996) tries to solve the overlapping routes problem

(see Figure 2.5) by considering an additional attribute which describes the commonality of routes. If a route is constructed of edges exclusively belonging to this route, this commonality factor equals to zero but the larger the overlap between paths, the larger their commonality and the smaller their utility. There are different formulations of the commonality factor expressing different concepts of similarity[14] but the C-Logit model's major disadvantage is that it is still not able to capture all aspects of similarity and that some formulations of the commonality factor result in contra-intuitive outcomes (Prato 2009).

Like the C-Logit model, the Path-Size Logit (PL) model (Ben-Akiva and Bierlaire 1999) adds a correction term in order to deal with overlapping routes (see Figure 2.5) but it has a different theoretical basis and interprets route similarity in a different way using the ratio between edge and route length. Because the original formulation was not suitable for long trips, it has been generalised (Ramming 2002) and expanded (Frejinger, Bierlaire, and Ben-Akiva 2009). Running the PL model in its generalised form highly increases computational costs (Ben-Akiva, Ramming, and Bekhor 2004; Prato 2009) but in general, the PL model outperforms the C-Logit model (Ramming 2002; Prato 2009).

The Paired Combinatorial Logit (PCL) model (Chu 1989; Koppelman and Wen 2000) is a conceptual generalisation of the NL model allowing differential correlation between each pair of alternative routes during the pairwise choosing of alternatives. Similar to other modelling approaches, the overlapping routes problem (see Figure 2.5) is addressed by including a similarity index (Ben-Akiva and Bierlaire 2003; Prato 2009).

2.4.2.3 Traffic assignment

Traffic assignment is an alternative approach to model route choice. It is based on the assumption that traffic flows in a street network are the sum of the route choices of all traffic participants. Traffic assignment aims at determining these flows and the resulting traffic conditions by making assumptions about travellers' route choice behaviour, modelling their choices, and representing the flows of traffic in the network (Chiu et al. 2011). A common behavioural assumption is that traffic participants chose the fastest available route between their origin and destination which ultimately leads to a user equilibrium in which no traveller can optimise their route choice (see Subsubsection 2.4.2.1 and Figure 2.4a). Based on these assumptions, traffic assignment algorithms try to calculate travel times and the traffic volume of each edge in the network (Chiu et al. 2011). However, representing traffic is complex: Historically, aggregated measures over rather long

[14] The commonality factor might depend exclusively on the distance shared between routes or it can be calculated using weights on the edge importance and ratios between edge and route lengths (Prato 2009).

time periods were used assuming that traffic is stable over the analysis period (Chiu et al. 2011). This so-called Static Traffic Assignment (STA) is not suitable for many applications because it ignores the time-dependent variation of traffic load (Jayakrishnan, Tsai, and Chen 1995; Zou et al. 2013; Krishna, Katti, and Gaurang 2015). Static models are also not able to represent the spill-back of congestion (Chiu et al. 2011). Dynamic Traffic Assignment (DTA) therefore aims at representing time-dependent variations in traffic conditions and flows. This requires an adjustment of the presumed equilibrium in that it is assumed that traffic participants 1) learn from past journeys and anticipate future traffic conditions and 2) minimise not theoretical but actual travel times. This in turn requires a dynamic representation of traffic flows and conditions as travellers experience different traffic conditions and travel times when they travel between the same origin-destination pair but at different times (Chiu et al. 2011). DTA can assist route choice by providing real-time traffic information and has therefore gained increased attention in traffic management, emergency planning, and traveller information science (Li et al. 2013).

2.4.2.4 Machine learning

Since ML was developed as a field of research in the early 1980s, it has grown steadily and the range of ML techniques has broadened and pushed the boundaries of computer science (Sammut and Webb 2011; Domingos 2012). Although advanced ML methods, such as random forests, Support Vector Machines (SVMs), and artificial neural networks, were successfully used in analysing consumers' behaviour (Kruppa et al. 2013; Gupta and Pathak 2014), traffic speed prediction (Tang et al. 2017), and transportation mode choice behaviour (Jahangiri and Rakha 2015), relatively little effort was made to apply these algorithms to the analysis of route choice behaviour (Lai et al. 2019). The remainder of this paragraph therefore briefly introduces some ML methods which have been applied in researching route choice behaviour.

An early ML technique which has also been used in route choice analysis is the use of decision trees (Fürnkranz 2011). A decision tree is an easily understandable tree-structured classification model to solve decision problems. A decision tree consist of a root, nodes and leaves whereby nodes represent rules and leaves represent answers. However, Lai et al. (2019) argue that, although the technique has been used in route choice analysis, decision trees are not robust enough to capture all its aspects and easily outperformed by other methods.

Reinforcement learning is a form of ML which is based on training through repetitive simulation. By trial and error, an agent discovers which actions are most promising and lead to a reward and which actions do not bring any benefit. Reinforcement learning

is widely considered as a promising approach in building artificial general intelligence (Sutton and Barto 2018). Miyagi (2004) used an adapted model to determine the user equilibrium in a crowded traffic network which in turn can be used to determine the interactions between driver's choices and the resulting user equilibrium. Another application is presented by Wei, Ma, and Jia (2014) who modelled drivers' experience and learning process. Both studies concluded that their approaches are promising, however, Wei, Ma, and Jia (2014) also emphasised the need for empirical studies to further investigate route choice behaviour.

A comparison between a reinforcement learning approach and another type of ML algorithms was presented by Sun and Park (2017) who predicted the route choice decisions of 18 participants with a neural network and with a SVM whereby the SVM outperformed the neural network by achieving similar prediction accuracy with much lower computing cost. SVMs belong to the class of maximum margin models and try to separate classes by representing each object as a vector in a vector space and then fitting a hyperplane that separates classes with the widest margin possible (Zhang 2011). The method was also used by Barua (2019) to model route choice behaviour. However, the choice set used in his study only consisted of two alternative routes which is too small to be considered realistic.

Fuzzy logic is another popular field in ML. As the name already indicates, the approach aims at capturing the vagueness and inaccuracies which are inherent in many everyday situations. Fuzzy logic is based on so-called fuzzy sets in which the elements are not absolutely assigned to a set but their membership is described by a membership function allowing partial membership in multiple sets (Cintula, Fermüller, and Noguera 2017).[15] Fuzzy logic has been used to model uncertainty in route choice behaviour (Hawas 2002; Henn 2003; Murat and Uludag 2008; Dhulipala et al. 2017).

2.4.3 Current research

This subsection presents an overview of the state of the art in researching route choice behaviour and related fields. According to the topic of this thesis, the focus is on research using taxi FCD. More comprehensive overviews are presented in Morikawa et al. (2005) and Jing et al. (2018).

Taxi FCD was used to investigate the influence of the underlying street network on traffic by Jiang, Yin, and Zhao (2009). Based on over 72'000 trips in four Swedish towns, they illustrate that the street network's topology and the spatial distribution of origin and des-

[15] An element's membership of a given set is usually described by a real value from the interval [0,1] where 0 represents "entirely outside of the set", 1 represents "entirely included in the set" and any value in between represents an intermediate degree of membership (Cintula, Fermüller, and Noguera 2017).

tination locations strongly account for the mobility patterns while the purposive nature of the taxi trips only has marginal influence. Jiang, Yin, and Zhao (2009) support their findings by reproducing the human mobility pattern using random walk simulation.[16] A second study confirming the strong influence of city layout and street network on travel patters was conducted by Liu et al. (2015) who analysed over 800'000 taxi trajectories from Shanghai, China. They applied clustering and community detection methods on the street network to reveal a hierarchical polycentric city structure and found that the purposes of different areas, such as business or residence, were strongly influencing the taxi drivers' mobility patterns while the influence of administrative boundaries on the latter was negligible.

Another study from Shanghai was presented by Peng et al. (2012). Their work investigated over 1.5 million taxi trips regarding the purpose of the trips with the goal to assess traffic flow in different locations and on different workdays. They found that on workdays, taxi trips are made mainly for either commuting between home and work location, for travelling between workplaces, or for other activities. Peng et al. (2012) then modelled total traffic flow as a linear combination of these three main categories. According to the authors, urban traffic at any given location can be modelled as a combination of basis patterns. They argue that their method provides an economical approach to infer land use and traffic composition which in turn are contributing factors in drivers' route choice behaviour.

In many cities, taxi drivers actively look for new costumers by driving around resulting in a substantial amount of additional traffic (Liu, Andris, and Ratti 2010; Yuan et al. 2011). Addressing this problem, a probability-based recommending system for taxi drivers and passengers was proposed by Yuan et al. (2011). Their LBS system directs drivers to locations with a high probability of picking up passengers and provides searching passengers with nearby locations where they can find a taxi. The predictions are based on the passengers' mobility patterns and on drivers' pick-up locations which were extracted from GPS trajectories. Yuan et al. (2011) validate their model with a large set of taxi trajectories concluding that their method is accurate enough to lower the number of unoccupied taxi trips and can therefore reduce traffic and environmental pollution. Furthermore, their research strongly suggests that finding a new passenger is the driving factor of taxi drivers' route choice behaviour during unoccupied trips.

One of the first large-scale studies using FCD to directly investigate taxi drivers' route choice behaviour was conducted by Liu, Andris, and Ratti (2010) and involved over 48 million trips from Shenzhen, China. Based on daily income, they categorized the drivers

[16] Random walks are models of simple stochastic processes and often used to model randomness in networks (Masuda, Porter, and Lambiotte 2017).

into top earners and ordinary drivers and analysed the trajectories in terms of operation tactics, spatial selection behaviour, context-aware spatio-temporal operation behaviour, and route choice behaviour. The latter was analysed by comparing the lengths and travel times of the actually driven routes with the corresponding shortest routes. Liu, Andris, and Ratti (2010) found that in general, taxi drivers are not trying to optimise trip length but try to finish their trip as efficiently as possible. Top earning drivers tended to drive faster but also showed a better knowledge of the street network and the traffic conditions which gave them more flexibility during their trips. Another surprising finding of Liu, Andris, and Ratti (2010) was that high-earning drivers did not make their profit in the city centre but in other parts of the city. A study conducted by Yao et al. (2013) presented similar findings based on taxi FCD from Beijing, China. The authors used an MNL model (see Subsubsection 2.4.2.2) to calculate the choice probability of different routes whereby increasing a route's number of left turns or its proportion of minor roads had the same effect on the results as decreasing the route's travel speed. From this, Yao et al. (2013) concluded that taxi drivers tend to optimise their routes for speed by choosing routes with less left turns and a high proportion of major roads and express ways. Another work using Beijing taxi data was presented by Li, Wang, and Wang (2018) who analysed the trajectories regarding cost-based route choice rules, namely shortest distance, shortest time, lowest number of signalised intersections, and lowest number of turns. They found that more than half of the actual route choices could not be explained based on these rules. However, assuming that the taxi drivers do not exactly follow an optimal route, but the route they choose is very similar, over 90 % of their route choice could be explained. From this, Li, Wang, and Wang (2018) concluded that taxi drivers take an imperfect but satisfactory route.

While the previously mentioned work focused on trips with passengers on board, Cai et al. (2016) used over 2 million occupied and unoccupied trips from Beijing, China for an analysis based on the trips' length, spatial coverage, and spatial distribution. Their results indicate that taxi drivers' travel patterns show similarities but also difference with individuals' driving characteristics. Cai et al. (2016) further argue that occupied taxi trips do not accurately represent individual vehicular travel and that an analysis should therefore include trips without passengers to correctly assess traffic participants' route choice behaviour. However, this conclusion may be questioned as taxi drivers in search of passengers do not display target-oriented driving behaviour but mainly cruise arterial roads (Morikawa et al. 2005; Nian, Zhu, and Sun 2017). A study analysing the differences in route choice between occupied and unoccupied taxi trips was presented by Nian, Zhu, and Sun (2017) who analysed 1'200 trips from Shenzhen, China. They state that driving characteristics between occupied and unoccupied taxis show large dif-

ferences and that in general, drivers prefer arterial roads and try to minimise the number of signals, expected travel time, congestion while unoccupied drivers seem to care less for signals count and travel time. Another Study utilising data from Shenzhen was presented by Sun et al. (2014) who compared taxi drivers' routes with shortest and fastest routes. They concluded that although only few drivers follow the shortest or fastest route, travel distance and travel time, as well as road preference, have relatively high influence on the taxi drivers' route choice.
One of rather few studies using European taxi FCD was presented by Manley, Addison, and Cheng (2015). The authors empirically analysed 700'000 trips in London with regard to the influence of certain urban features on route choice. They argue that the basis of route choice are major urban features because they attract more route choices than cost minimisation could explain. Furthermore, Manley, Addison, and Cheng (2015) reinforce the argument that shortest distance is not able to accurately predict actual routes.

2.4.4 Route similarity

Much of the research presented in Subsection 2.4.3 compares taxi drivers' route choices with alternative routes. However, the extent to which similarity between different routes is described varies. Yao et al. (2013), for example, only distinguished between completely congruent routes and not completely congruent routes without any gradation of route similarity while Liu, Andris, and Ratti (2010), Sun et al. (2014), and Manley, Addison, and Cheng (2015) used indices and ratios based on route characteristics, such as length or duration, which recorded not only whether, but also to what extent, two routes are similar. More complex approaches were used by Li, Wang, and Wang (2018), who calculated so-called coverage and performance scores to compare different routes, and Cai et al. (2016) who proposed a method which considers the routes' spatial coverage, visited locations, sampling rate, and more. In this thesis, route similarity will be accessed using multiple measures which are described in the remainder of this subsection.

2.4.4.1 Percentage of shared length

The Percentage of Shared Length (PSL) (Sun et al. 2014) describes how much of its length an actual route shares with an optimal route[17]. It is calculated using:

$$PSL = \frac{L_{shared}}{L_{actual}} * 100, \tag{2.2}$$

where L_{shared} is the accumulated length of the street segments travelled by both routes

[17] The term "optimal routes" refers to shortest, fastest, and fewest intersections routes.

and L_{actual} is the length of the actual route. The result is a percentage value which refers to the actual route as 100 %. Calculating the PSL requires knowledge of which street network edges are traversed by both routes as well as of the individual edges' lengths.

2.4.4.2 Hausdorff distance and Fréchet distance

Since the PSL is based on shared street network edges, it only discovers route similarity if two routes are both using exactly the same path. However, assuming that a driver chooses a route which travels a street running parallel to the optimal route, no similarity will be detected. This problem is addressed by geometric approaches which are not depending on the underlying street network but only consider geometrical aspects (Feld 2019). This paragraph briefly introduces two techniques to measure curve similarity, namely the Hausdorff Distance (HD) (Hausdorff 1927) and the Fréchet Distance (FD) (Fréchet 1906), and refers to Alt and Guibas (2000) and Feld (2019) for a more comprehensive overview.

The HD is a well known similarity measure for the comparison of shapes and patterns and constitutes the basis for many related distance functions (Alt and Guibas 2000). It interprets shapes as sets of points and assigns each point in a set with the distance to its closest point in another set. The result is then the maximum of these shortest distances whereby small values indicate a high similarity between the two sets. In the context of spatial trajectories, the HD is the maximum of shortest distances between two trajectories (Alt and Guibas 2000; Feld 2019). A disadvantage of the HD is its high susceptibility to noise, as a single outlier can significantly affect the final result (Veltkamp 2001), and therefore, this thesis uses the more complex and more robust FD.

The FD is a measure of similarity between curves and an extension of the HD which not only considers the points locations, but also their order (Alt and Guibas 2000; Feld 2019). It is defined as the minimum length of a link connecting two points travelling with varying speeds on two different trajectories whereby a small HD indicates high similarity between the trajectories. The concept is usually illustrated by a man walking his dog. Both, the man and the dog, are allowed to vary their speed or to stand still but neither of them can move backwards. In this scenario, the FD is the shortest possible length of the leash required (Feld 2019) (see Figure 2.6). Calculating the exact FD is computationally very intensive (Feld 2019) but the so-called discrete FD (Eiter and Mannila 1994) provides an approximation which can be computed in polynomial time by a relatively simple algorithm. This is achieved by only considering positions located on the trajectories' vertices for the calculation of the leash's length but never positions located in the interior of an edge (Feld 2019) (see Figure 2.6).

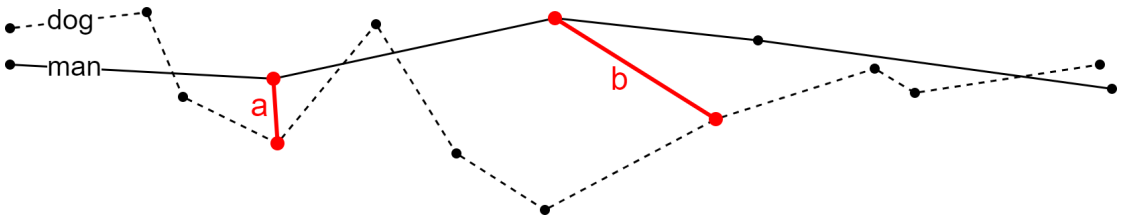


Figure 2.6: The two lines represent the two trajectories traversed by a man and his dog with a leash (red) between them. After starting the walk, they are close to each other (a) but then, the dog walks much faster than the man increasing the length of the leash (b). Figure modified from Van Diggelen (2018, p.3).

2.4.4.3 Percentage difference of length, duration, and number of intersections

The Percentage of Length Difference (PLD), Percentage of Time Difference (PTD), and Percentage of Intersections Difference (PID) are three alike indices describing how an actual route differs from an optimal route in terms of length, duration, and number of intersections. The indices are calculated using:

$$PLD = \frac{L_{actual} - L_{optimal}}{L_{optimal}} * 100, \tag{2.3}$$

where L_{actual} is the length of the actual route and $L_{optimal}$ is the length of the optimal route,

$$PTD = \frac{T_{actual} - T_{optimal}}{T_{optimal}} * 100, \tag{2.4}$$

where T_{actual} is the optimal duration of the actual route and $T_{optimal}$ is the optimal duration of the optimal route, and

$$PID = \frac{I_{actual} - I_{optimal}}{I_{optimal}} * 100, \tag{2.5}$$

where I_{actual} is the number of intersections in the actual route and $I_{optimal}$ is the number of intersections in the optimal route. The result of all three formulas is a percentage value which refers to the optimal route as 100 %, meaning that positive values indicate the actual route being longer, slower, or visiting more intersections than the optimal alternative. In contrast to the PSL, calculating the PLD, PTD, and PID does not require knowledge of partial routes or street network edges but only information about

the routes' total length, duration, or number of intersections.

2.4.5 Research gaps

Yao et al. (2013), Sun et al. (2014), and Manley, Addison, and Cheng (2015) have presented the key studies regarding the route choice behaviour of taxi drivers with passengers on board. However, there is room for further research in this area as their work presents the following research gaps:

- The studies use only data from a single city, namely Beijing (Yao et al. 2013), Shenzhen (Sun et al. 2014), and London (Manley, Addison, and Cheng 2015).
- The number of analysed trips is relatively small, namely 211 (Yao et al. 2013), 40'000 (Sun et al. 2014), and 680'000 (Manley, Addison, and Cheng 2015).
- The number of intersections has not yet been investigated as a potential influencing factor of taxi drivers' route choice.

Addressing these research gaps, the analysis presented in this thesis is based on data from three different cities and investigates a total of over 2.1 million taxi trips. Furthermore, it compares taxi drivers' routes not only with shortest and fastest but also with fewest intersections routes.

Chapter 3

Data and study areas

3.1 Street network data

The networks used for MM and shortest path computation were retrieved from OpenStreetMap (OSM) (OpenStreetMap contributors 2017) and reflect the current state of the cities' street network.[19] OSM is a crowd-based map database with global coverage which is built, maintained, and expanded by volunteers. The project's open data policy and its data access APIs allow easy download and use of map data (OpenStreetMap contributors 2018, 2019). Table 3.1 provides an overview of the networks which vary considerably in size and complexity. For a visual comparison of the networks, see Figures 3.1, 3.2, and 3.3.

Table 3.1: Overview of the street networks and the resulting graphs.

	San Francisco	**Shanghai**	**Vienna**
Total street length (km)	1'869	24'690	3'028
Oneway street length (km)	594	10'298	1'456
Oneway streets (%)	31.8	41.7	48.1
Street density (km/km^2)	12.0	2.4	7.0
Intersections	8'528	112'501	20'693
Intersections per km^2	54.7	11.1	47.7
Average edge betweenness centrality	0.00166	0.00072	0.00201
Nodes in graph	60'017	271'421	81'465
Edges in graph	106'969	466'769	133'923
Graph convex hull (km^2)	156	10'107	434

3.2 Floating car data

The FCD originate from multiple sources: the San Francisco dataset was provided by Piórkowski, Sarafijanovoc-Djukic, and Grossglauser (2009). It was collected in May and June 2008 and contains over 11 million data points representing trips from 536 individual taxi drivers. The Shanghai data was shared by Prof. Chun Liu from Tongji University in Shanghai and contains almost 700 million data points collected by 7'758 drivers during a two-week period in June 2010. The third dataset, containing 54 million data points, was collected in Vienna between June and November 2015 and was provided by the Austrian Institute of Technology in Vienna which received the the data from a private taxi company.[18] As Table 3.2 shows, the datasets vary in number of records, temporal resolution, time period[19], and included attributes.

Table 3.2: Overview of the FCD.

	San Francisco	**Shanghai**	**Vienna**
Data points	11'219'955	698'852'167	54'093'068
Unique drivers	536	7'758	unknown
Temporal resolution	∼60 seconds	∼10 seconds	∼40 seconds
Time period	May – Jun 2008	Jun 2010	Jun – Nov 2015
File format	.txt	.txt	.txt
File size	361 megabytes	55.7 gigabytes	8.7 gigabytes
Attributes	driver id*	taxi id*	trip id*
	latitude*	latitude*	latitude*
	longitude*	longitude*	longitude*
	occupancy*	occupancy*	occupancy
	timestamp*	timestamp*	timestamp*
	updates	date	heading
		velocity	covered distance
		direction	geometry
		...	...

* attributes used for pre-processing (see Table 4.2)

[18] For further information about the taxi company, see https://www.taxi31300.at/.
[19] Note that the street network data does not represent the street networks during the time of the FCD collection.

3.3 Study areas

3.3.1 San Francisco

The city of San Francisco is located on the US West Coast on a peninsula between San Francisco Bay and the Pacific Ocean. It occupies a a hilly area of about 120 square kilometres and is surrounded by water on three sides with two bridges connecting the city to the mainland in the north and east. As a cultural and financial centre, it is one of the most cosmopolitan cities in the country with a population of 800'000 in July 2008 (State of California 2011; Conrad et al. 2019). Because San Francisco's street network is arranged in a rectangular grid, some of the roads have long and steep slopes (Conrad et al. 2019). Figure 3.1 shows the study area used in this thesis and the city's network of drivable streets.

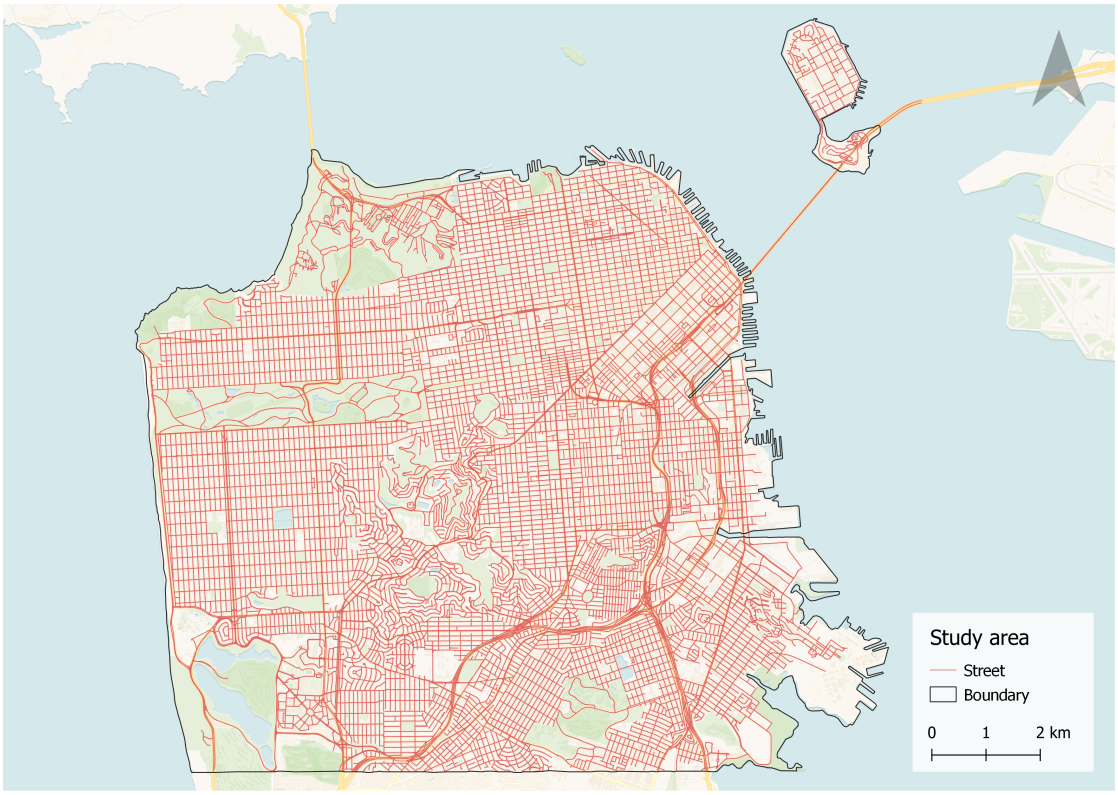


Figure 3.1: Study area and drivable street network in San Francisco.

3.3.2 Shanghai

Shanghai is China's most populated city, with over 20 million inhabitants in 2010, and one of the country's major industrial and commercial centres. It is located on the Chinese East Coast in the Yangtze River Delta and occupies an area of 6'340 square kilometres. The mostly flat terrain is crossed by the Huangpu River and an extensive network of smaller waterways and canals which are heavily used for transportation. Despite the construction of express highways starting in the late 20$^{\text{th}}$ century, Shanghai's street network is frequently overloaded resulting in traffic jams and delays (Boxer 2019). The city's drivable streets, which were used in this analysis, are shown in Figure 3.2.

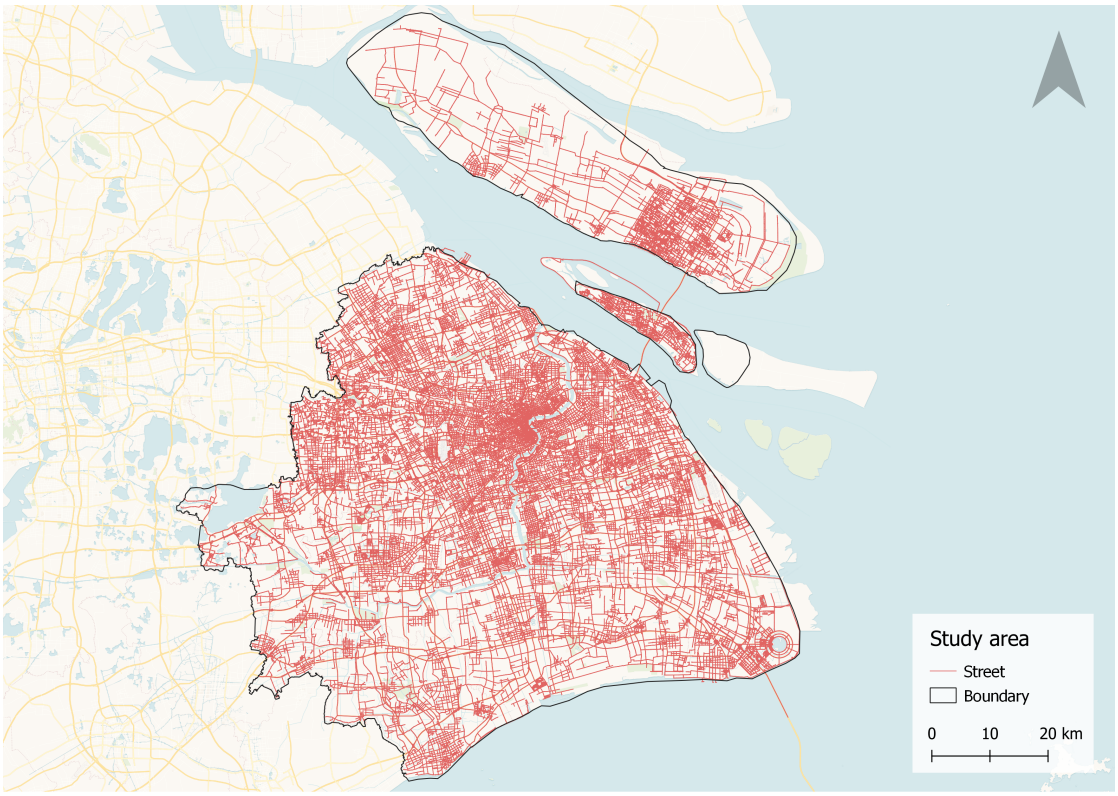


Figure 3.2: Study area and drivable street network in Shanghai.

3.3.3 Vienna

The city of Vienna lies on the banks of the Danube River in the north-east of Austria between the Alps and the Carpathians. The city is the capital of Austria and the country's economic and cultural centre with a population of 1.8 million in 2015 (Statistical Office of the City of Vienna 2015). Occupying an area of 415 square kilometres, Vienna is situated on multiple terraces which vary considerably in height. Regarding mobility, most people use the modern and extensive public transportation network for their travels (Ehrlich et al. 2020). The street network contains multiple motorway axes and about 1'700 bridges (City of Vienna 2020). The drivable streets and the study area are shown in Figure 3.3.

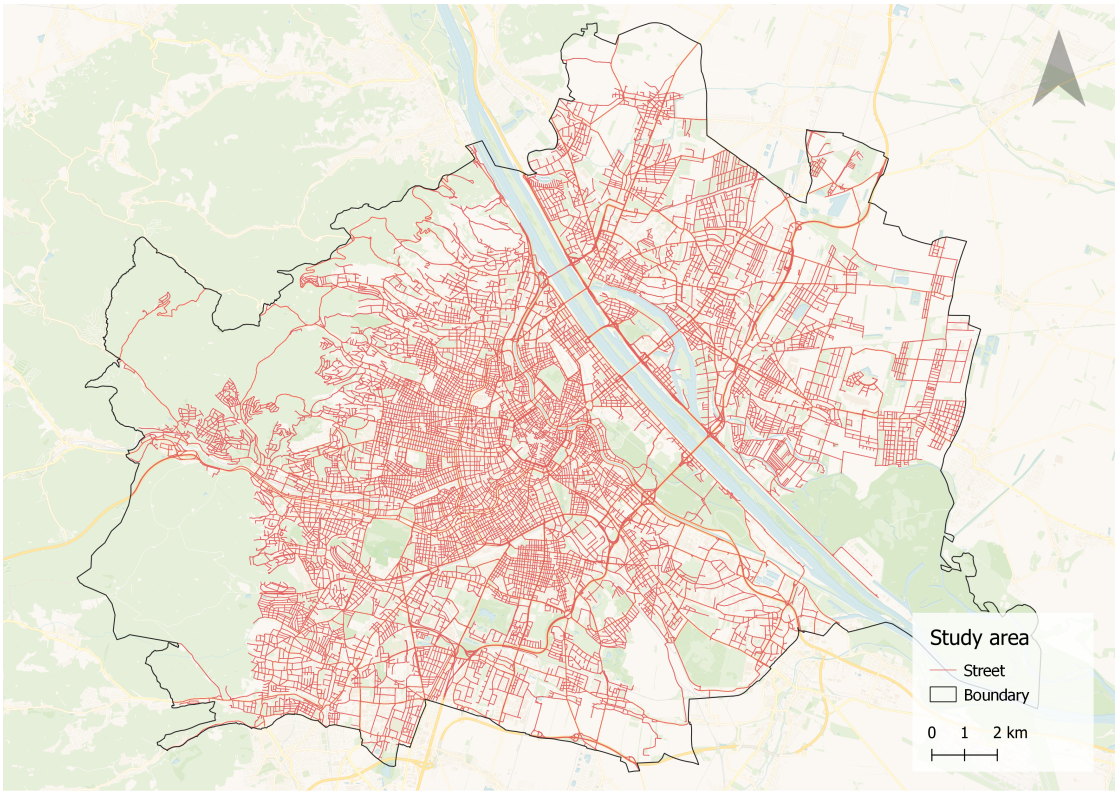


Figure 3.3: Study area and drivable street network in Vienna.

Chapter 4

Methodological approach

4.1 General workflow

The general workflow of all data processing conducted for this thesis is presented in Figure 4.1. The flowchart contains all major processing steps for the street network data and the FCD and provides an overview how the data were prepared for the analysis. Preparing the network data was the much smaller task and is described in Section 4.3. The three datasets were processed identically since they were identically structured for all three cities using predefined and custom functions in Python. As the FCD varies in size and complexity, the three datasets were brought into the same form so that they could be processed using the same functions and scripts. After the trajectories were pre-processed and the actual routes were derived (see Section 4.4), optimal routes were computed (see Section 4.5) and post-processed together with the actual routes (see Section 4.6) before they were finally merged to the final routes dataset (see Section 5.1) providing the basis for the analysis. The following section presents the various tools used for handling the data while the remaining chapter describes the individual processing steps shown in Figure 4.1.

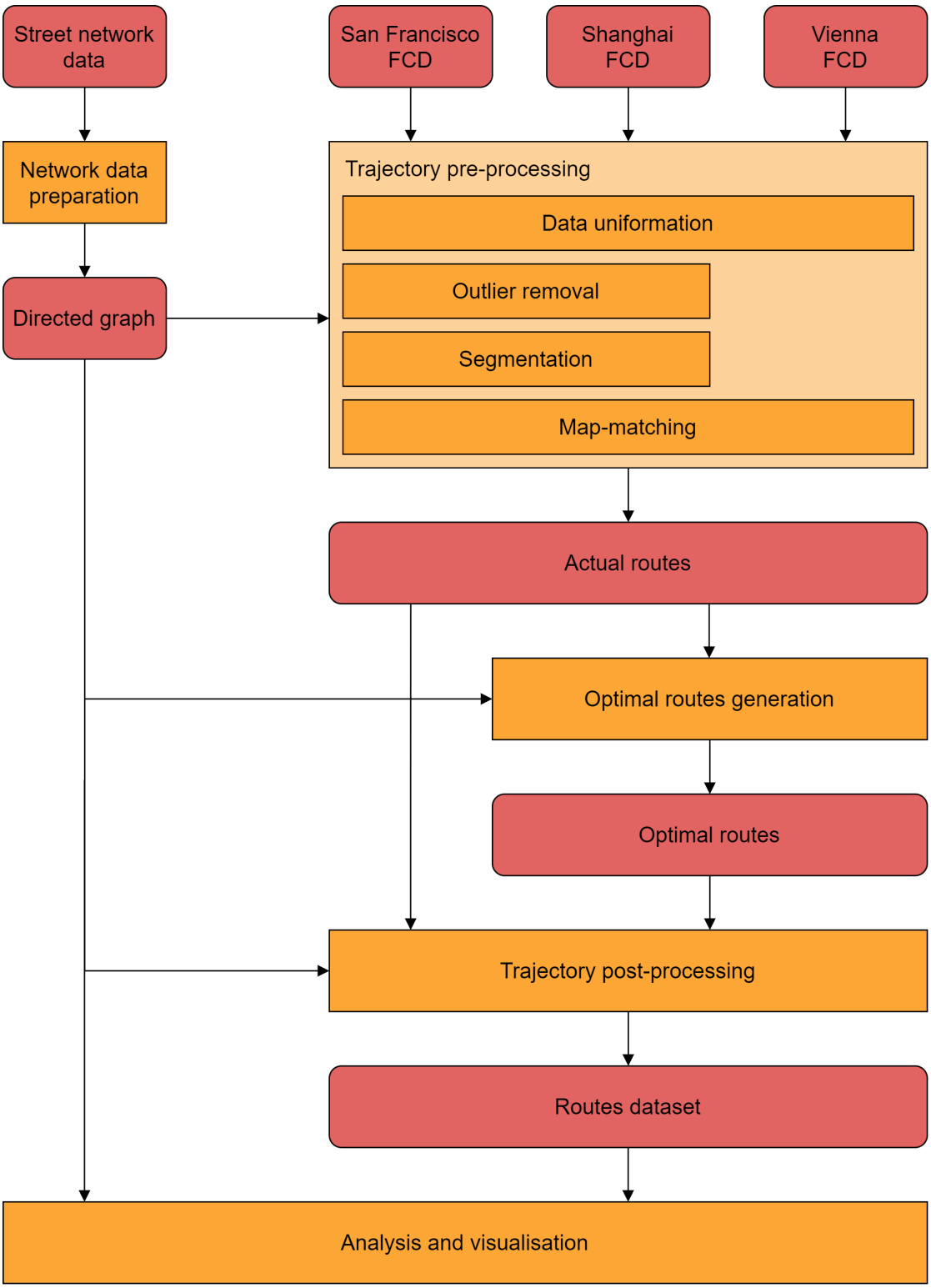


Figure 4.1: Worflow of data processing The processes (orange boxes) correspond with the chapter's sections and subsections. Red boxes represent data. Note that some pre-processing steps were left out for the Vienna data since they were already done previously to this thesis (see Section 4.4).

4.2 Tools

4.2.1 Python and R

Python (Van Rossum and Drake 2009) is a powerful high-level scripting language which has established itself as a primary tool for scientific computing over the last decades. Its biggest strengths are its user-friendly syntax and the large ecosystem of freely available packages extending the core functionalities (VanderPlas 2016). Python[20] has been the language of choice for most of the programming throughout this thesis using PyCharm as an IDE. The most used packages were:

- pandas (McKinney 2011) and NumPy (Oliphant 2006) for data manipulation.
- OSMnx (Boeing 2017), NetworkX (Hagberg, Schult, and Swart 2008), and graph-tool (Peixoto 2014) for network manipulation and shortest path search.
- GeoPandas (GeoPandas Development Team 2020) and Shapely (Gillies et al. 2007) for spatial data manipulation and analysis.
- SQLAlchemy (Bayer 2012) for interacting with the PostgreSQL database.

Besides Python, the programming language R[21] (R Core Team 2019) was used for creating plots and non-spatial graphics. Like Python, R benefits from a variety of third-party libraries providing additional functionalities. For this thesis, the two most frequently used libraries were dplyr (Wickham et al. 2019) for data manipulation and ggplot2 (Wickham 2016) for creating plots and graphics. RStudio was used as an IDE.

4.2.2 PostgreSQL

In addition to the raw data described in Chapter 3, another 400 gigabytes of data were generated in the course of this work. This amount of data required to be handled very efficiently and therefore, all data generated after map matching were stored and managed using PostgreSQL[22] (The PostgreSQL Global Development Group 2020) which is a free and open-source object-relational Database Management System (DBMS). Its functionalities were further extended by PostGIS[23] providing support for spatial and geographic data. The database contained each cities' street network together with the actual and optimal routes as well as route characteristics and some additional attribute tables. In

[20] versions 3.6.1 and 3.7.5
[21] version 3.5.3
[22] version 11.5
[23] For further information, visit `https://postgis.net/`.

total, about 290 gigabytes of data were stored in 34 tables. The remaining 110 gigabytes, namely the data generated before MM, were stored on a conventional hard disk to reduce the costs for the database. The database was run using Amazon RDS.[24] The administration tool pgAdmin[25] was used to interact with the database.

4.2.3 QGIS

All spatial visualisations were created using QGIS[26] (QGIS Development Team 2020). This Geographic Information System (GIS) is open-source and offers a variety of functionalities for editing, analysing, and visualising spatial data. Its capabilities can be further extended by external plugins. In addition to visualisation, QGIS was also used to explore the datasets and preliminary results using the possibility to directly connect QGIS to a PostGIS database.

4.2.4 Hardware

All computations were done on two standard systems running on Microsoft Windows 10: one with an 8-core processor at 4 gigahertz and 16 gigabytes of RAM and one with a 4-core processor at 2.5 gigahertz and 8 gigabytes of RAM. The use of external servers or cloud computing solutions was deliberately avoided in order to maintain the greatest possible control over the software used. However, this also meant that CPU-intensive tasks needed to be performed as efficient as possible, which was achieved by consequently using vectorised functions instead of loops and by outsourcing shortest path searches to C++ (see Section 4.5). In order to deal with the limited memory, the data was mostly processed in subsets.

4.3 Network data preparation

The network data has been retrieved from OSM using the OSMnx (Boeing 2017) module for Python. This package not only allows the extraction of street networks from OSM's APIs, but offers a variety of functions to edit, model, analyse, visualise, and save them.[27] It also gives the option to reduce a graph's complexity by simplifying its topology. However, simplification is achieved by removing all nodes which are not dead-ends or intersections leading to the straightening of curves and therefore to distorted distances. For this reason, the unsimplified networks and resulting graphs described in Section 3.1

[24] For further information, visit https://docs.aws.amazon.com/AmazonRDS/latest/UserGuide/Welcome.html.
[25] For further information, visit https://www.pgadmin.org/.
[26] version 3.8.2
[27] For further information, visit https://geoffboeing.com/2016/11/osmnx-python-street-networks/.

were used during all stages of this thesis.
Using OSMnx, the drivable street network from each city was downloaded from OSM and uploaded into a PostGIS database. At this stage, the networks were represented as undirected graphs, defined by "from"- and "to"-relations in edge lists referring to the corresponding node list. Additionally, the attributes extracted from OSM were stored. Using the attribute determining each edge as one way or two way street, all edges representing two way streets were duplicated with reversed start and end nodes in order to construct directed graphs accounting for mobility restrictions in the street networks. This means that each bidirectional street segment is stored as two opposite directed edges. Further, the large integer values of the OSM node ids were relabelled into smaller values to prevent numeric problems during the later MM process.
As preparation for the trajectory post-processing, the normalised EBC (see Subsubsection 2.1.3.2) was computed for all three graphs using Brandes' algorithm (Brandes 2001) and the graph-tool (Peixoto 2014) module for Python. The values were then stored as an additional edge attribute. Furthermore, each edge was assigned the maximum speed of the corresponding street segment. The OSM data already included a speed limit attribute but the values were incomplete and missing for large parts of the cities. The maximum speed values were therefore assigned based on the road types which are also stored in the OSM data (see Table 4.1). The limits were chosen based on Beijing Expat Service Center (2019) and OpenStreetMap contributors (2020c,b,a) but since OSM's data are crowd-sourced, the data are somewhat subjective.[28] However, the values within each city are consistent and therefore allow at least city-wide comparison. If the source suggested several different speed limits for a road type, the lowest value was chosen. Finally, the travel time for each edge was calculated based on its length and speed limit.

[28] An example for this are the presumably high speed limits for residential streets (see Table 4.1) which have been proposed by OSM.

Table 4.1: OSM road types and speed limits.

Road type	Speed limit (km/h)		
	San Francisco	Shanghai	Vienna
Living street	5.0	5.0	5.0
Residential	40.0	30.0	40.0
Unclassified	40.0	30.0	50.0
Tertiary	40.0	30.0	50.0
Secondary	56.5	30.0	50.0
Primary	56.5	30.0	50.0
Trunk	72.5	100.0	100.0
Motorway	104.5	120.0	130.0
Other	40.0	30.0	50.0

4.4 Trajectory pre-processing

4.4.1 Data uniformation

The FCD from the three cities varies in its structure and size (see Table 3.2). In a first step, the three datasets were therefore brought into a similar structure so that each data point was attributed with its time and location of recording as well as information about the driver or trip and whether a passenger was on-board (see Table 4.2). It is worth mentioning that the data points from Vienna were already assigned to trips whilst the data from San Francisco and Shanghai did not contain this information. Further, the Vienna data had already been map matched and trips without passengers were removed so that, apart from restructuring the data, no further pre-processing steps were required for this particular dataset (see Figure 4.1).

Since the temporal resolution of the Shanghai raw data is 4–6 times higher than in the other two datasets, only every fourth data point has been used for further pre-processing, resulting in a temporal resolution of about 40 seconds which is similar to the Vienna data (see Table 3.2). This was done because it makes the different datasets more similar and therefore reduces the influence of using different data sources on the results. Furthermore, this step significantly reduces the number of records which is beneficial for system workload and running times. It is assumed that the quality of the MM results do not suffer from a lower temporal resolution since a time interval of 1 minute between

consecutive data points is sufficient for the chosen MM algorithm (see Subsection 4.4.4) (Yang and Gidófalvi 2018). After equalising the data, the following steps were performed for the San Francisco and Shanghai datasets.

Table 4.2: Data attributes used for pre-processing.

Field	**Type**	**Comment**
driver*	Integer or character	Unique identifier for each driver
id**	Integer	Unique identifier for each trip
time	Integer	Unix timestamp
lon	Float	Longitude (WGS84)
lat	Float	Latitude (WGS84)
occ*	Integer	1 if passengers on board, else 0

* only in the San Francisco and Shanghai data

** only in the Vienna data

4.4.2 Outlier removal

Since this thesis focuses on taxi drivers' route choice behaviour in urban areas, all data points outside of the cities' boundaries were removed. Furthermore, all data points which were not located within 20 metres of a street were removed by spatially intersecting the points' locations with a buffer polygon constructed from the street network. This particular threshold was chosen because the upper limit of GNSS positioning errors in urban areas lies somewhere between 20 and 30 metres (Ververidis and Polyzos 2006; Li et al. 2018). Points with a speed value above 180 kilometres per hour were also removed. Because speeds were calculated based on distance and time difference between consecutive data points, the resulting values are affected by GNSS positioning inaccuracies. The threshold to identify these speed outliers was therefore chosen rather tolerant.

4.4.3 Segmentation

Stops were detected when either the driver or the occupancy changed. Then, trips with no passengers on-board were removed. After recalculating distance and time difference between consecutive points as well as speeds in the cleaned data, stops were detected when the average speed of a point was lower than 0.5 kilometres per hour. In order to detect only long stops, a moving window, considering 10 consecutive data points, was used to calculate the speed values. Finally, unique trip ids were assigned after each detected

stop, after the driver changes, after a new passenger gets in, and after stops of about 4–6 minutes. At this stage, the three datasets were structured identically containing information about position, time, and trip for each data point. All remaining steps in this analysis were therefore performed identically for all three cities.
Using only trips with at least 5 data points, a LineString geometry was created for each trip and attribute with the respective trip id, start time, and end time. The trips were then written to multiple ESRI shapefiles as this was the preferred input format for the MM algorithm described in Subsection 4.4.4.

4.4.4 Map matching

In total, there were over 6 million trips, consisting of over 100 million points, which needed to be map matched. Additionally, the networks chosen for map matching were rather complex with up to over 460'000 edges in the Shanghai graph (see Table 3.1). Since the computational resources for this thesis were limited (see Subsection 4.2.4), choosing an efficient MM approach was of particular importance. The requirements regarding performance and quality of results were met by the FMM algorithm presented in Yang and Gidófalvi (2018) (see Subsection 2.3.1).
The pre-processed trajectories and street networks from San Francisco, Shanghai, and Vienna (see Section 4.4) were map matched in 18 individual subsets. Although the Vienna data had already been map matched previously and were therefore not pre-processed along with the San Francisco and the Shanghai data, the map matching was deliberately redone in order to handle all three datasets identically, hoping to make the results more comparable. FMM presents multiple options regarding the format of the input data and the interface used to set the algorithm's parameters. In this thesis, the procedure proposed by FMM's documentation was applied which means that the use of locally stored ESRI shapefiles and a simple command-line interpreter was preferred over using the database and FMM's Python API although this would have been more convenient as less steps would have been required for the whole MM process. However, it was assumed that using the Python API would also make the whole process slower. After a first run of MM, about 10 % of the matched trajectories displayed unnatural routes including repeated reverse movement. Therefore, the whole process was redone with an increased penalty for reversed movement (see Subsection 2.3.1) which seemed to solve the problem since no such patterns were discovered. For a brief discussion of the MM results, see Section 5.1.

4.5 Optimal routes generation

After map matching the actual routes, three alternatives were computed for each of them: the shortest route, the fastest route, and the route with the fewest intersections. This was done by extracting the nodes closest to each actual route's start and endpoints from the prepared street network graphs (see Section 4.3) and searching for the connecting routes which fit the criteria best. By weighting each edge in the street network graph with its length, travel time, and number of intersections, this becomes a single-source shortest path problem (Ahuja, Magnanti, and Orlin 1993) which can be solved using the appropriate algorithm (see Subsection 2.1.4). Since the street network graphs are directed and only contain positive edge weights, shortest paths can be found using Dijkstra's algorithm (Dijkstra 1959).

The calculation of the shortest paths was the computationally most complex part of the thesis and given the large number of paths and the large size of the street network graphs, it had to be as efficient as possible. With the NetworkX (Hagberg, Schult, and Swart 2008) Python module, which was used during previous steps, it takes 0.1–0.15 seconds to compute a single shortest path in a comparable network which was not fast enough for the given task since it would have taken 2–3 weeks to calculate the over 12.9 million paths. Luckily, there is a handful of other Python modules for working with graphs from which graph-tool (Peixoto 2014) showed the most promise. Graph-tool's biggest advantage is its high performance, which comes from the core algorithms and structures being written in parallel C++, resulting in shortest path computing times about 40–80 times lower than with NetworkX (Peixoto 2015; Lin 2019). However, graph-tool does not come with installation-ready files for Windows and was therefore installed an run in the Ubuntu userspace for Windows 10.

Since the networks exported from OSM only contained length as an edge attribute, travel time and number of intersections had to be calculated. The travel time for each edge was estimated using its length and road type which are included in the OSM data. For the computation of fewest intersections routes, the number of intersections was also stored as an edge attribute. Since the presence of intersections is clearly determined by a graph's nodes and not its edges, this required an additional step in which each edge was assigned the average degree of its source and target node. This method assumes that a driver always travels both nodes connected to an edge which is not entirely true since it does not apply for the first and last edge in a path. However, given that a single path can travel hundreds of edges, this seems negligible.[29] After constructing the graphs from the

[29] Note that the calculation of the number of intersections visited by a route, as it is described in Section 4.6, does not make this assumption since it is based on the degree of the nodes actually visited by the route.

edge and node lists stored in the database (see Section 4.3), calculating the shortest paths was straightforward since graph-tool provides a function applying Dijkstra's algorithm. The function takes a graph, source and target nodes, as well as edge weights as input and returns an edge list with all edges travelled by the shortest path. It is worth mentioning that the directional graph represented one way streets, while turn restrictions were not taken into account.

4.6 Trajectory post-processing

During the MM process, all previous attributes but a unique trip identifier assigned during pre-processing were dropped. The metadata stored before map matching were therefore joined with the matched geometries based on this trip id. All paths which could not have been map matched or only contained one data point were removed. Using the street network graphs, a number of attributes, namely origin and destination edges and nodes, length, actual and optimal duration, number of traffic lights, and beetweenness centrality, were calculated for each of the map matched actual route paths. The length was calculated by simply adding up the length of all edges travelled by a path. The actual duration was calculated from the timestamps of the origin and destination points and the optimal duration was computed using the edge travelling times which were calculated during network data preparation (see Section 4.3). This means that the optimal duration displays the time a vehicle would need to cover the given route if it could drive with the maximum speed allowed and without slowing down on intersections, traffic lights, or crossings. This is unrealistic and a fact one should be aware of when using this attribute for further calculations. The number of traffic lights was determined based on a corresponding attribute in the OSM data and the betweenness centrality was calculated by taking the average edge betweenness centrality of all edges visited by the path.
The attributes described above were also calculated for each of the optimal routes computed during the steps described in Section 4.5. In addition, the routes' spatial geometries were constructed from the edge lists, returned by the shortest path search algorithm, and stored as LineString geometries (see Figure 5.3). In order to compare the routes in terms of shared length, the distance an actual route shares with each of the corresponding optimal routes was identified. This was done by merging the edge lists of the routes and finding duplicated edges since a duplicated edge indicates that both routes travel through it. It was also determined which road types are covered by each route. Furthermore, the number of intersections, as well as their complexity, was assigned as additional route attributes. For this purpose, the graphs' node degree (see Subsubsection 2.1.3.1) was

calculated and then, the number of nodes with a certain degree travelled by each route were counted whereby only nodes with a degree higher than 2 were taken into account. Based on these measures, the PSL, FD, PLD, PTD, and PID were calculated (see Subsection 4.6.1). Finally, the number of total turns, as well as specific turn characteristics, were computed by calculating the angles between subsequent points in the LineString geometries. To avoid counting curves as turns, a change of direction had to be at least 45 degrees to be considered a turn and a sharp turn was recorded when directional changes were greater than 90 degrees (Meng et al. 2009; Xu, Luo, and Shao 2018) (see Figure 5.2). Finally, all post-processed trajectories were stored in a database table.

4.6.1 Calculation of route similarity measures

In order to compare actual and optimal routes, multiple measures of route similarity (see Subsection 2.4.4) were calculated for each route. For the calculation of the PSL, it had to be determined which street network edges were shared between the actual and the optimal routes. The shared links were obtained by identifying the edge ids which were included in both routes' edge lists. These ids were then replaced with the lengths of the corresponding edges and summed up.

Of the 5 route similarity measures, the FD was the most complicated to compute. Although several Python modules include functionalities for its calculation, none of them was efficient enough so this step was finally done in R using the kmlShape library (Genolini 2016), which provides a fast C-compiled function to calculate discrete FDs. The function uses coordinate lists as input which were extracted from the LineString geometries representing the routes.

The PLD, PTD, and PID were calculated using the routes' lengths, optimal durations, and numbers of intersections which were computed earlier (see Section 4.6).

4.7 Analysis and visualisation

The data obtained during the steps described in the last sections were stored as separate tables in the database. However, in order to make the data smaller, clearer, and more suitable for export, all relevant results have been condensed in a single table which includes all routes and their respective attributes and geometries.[30] Entries with unrealistic values[31] were dropped and only trips where the actual and all optimal routes are available were kept. Furthermore, routes of extremely long or short distance or duration

[30] A comprehensive overview of all attributes contained in the dataset is given in the Appendix.

[31] This refers to values which were out of range and therefore indicated some error in the data processing. Examples are routes with a PSL greater than 100 % or routes with a duration of only a few seconds.

were removed in order to minimise the influence of anomalies, outliers, and edge phenomena on the analysis. In total, the routes dataset contains over 10.4 million routes for over 2.6 million unique trips in San Francisco, Shanghai, and Vienna (see Table 4.3).[32] Since the Vienna dataset contains routes from 6 months but the other two cities only have data from around June, all Viennese routes which were not conducted in June were removed. In the end, the data used for the results presented in Chapter 5 contained 8.5 million routes, which is 81.5 % of the whole routes dataset.
All plots presented in Chapters 5 and 6 were created in RStudio (see Subsection 4.2.1) for the reason that the R-universe incorporates the powerful ggplot2 library (Wickham 2016) and its extending packages for visualisation.

Table 4.3: Overview of the complete routes dataset. The term "Trip" refers to an actual journey between a given origin and destination without specifying the route that was taken while the term "route" refers to an actual or optimal route between a given pair of origin and destination.

Route type	**San Francisco**	**Shanghai**	**Vienna**	**Total**
Trips	170'544	1'451'476	990'544	2'612'564
Routes	682'176	5'805'904	3'962'176	10'450'256

[32] "Trip" refers to the journey between origin and destination without specifying which route is chosen. However, since the trips are derived from the origin and destination pairs, their number is equal to the number of actual routes. The term "route" refers to an actual or optimal route from an origin to a destination. Since four routes exist for each trip, there are four times more routes than trips.

Chapter 5

Results

5.1 Overview of the routes dataset

This section presents an overview of the routes dataset which was generated as basis for the analysis of taxi drivers' route choice behaviour. It addresses some key aspects such as Map Matching (MM) (see Subsection 4.4.4), optimal routes generation (see Section 4.5), and calculation of turns (see Section 4.6) and provides an overview of the data distribution in the dataset. A comprehensive list of all attributes included in the routes dataset can be found in the Appendix.

During map matching the GPS data points (see Subsection 4.4.4), the average matching percentage was 58.6 % with the best results in San Francisco where 88.8 % of trajectories were matched. However, percentage dropped to 73.0 % for Vienna and 53.3 % for Shanghai. These low matching percentages are likely due to the rather low search radius of 20 metres that was used and by increasing this parameter, more trajectories could have been mapped. However, since only data points within 20 metres distance to a street were kept during pre-processing (see Subsection 4.4.2), this might have led to wrong results. Increasing the number of points was no option since a buffer distance larger than the assumed GPS positioning error would suggest wrong relations between raw data points and streets. On visual inspection, the results seem to be of good quality (see Figure 5.1) and all things considered, it seems preferable to have a smaller set of trajectories, which are of high probability to be map matched correctly, than a larger but inaccurate dataset and therefore, this loss of information was accepted.
Since the number of turns was calculated based on the angles between adjacent points in the routes' LineString geometries without using information from the underlying street network (see Section 4.6), it was of particular importance to correctly distinguish between curves and turns (see Figure 5.2). The results of the turn characteristics calculation were

also assessed visually and seem to be satisfying as there were no apparent misclassifications in the data.

According to visual inspection, the post-processed optimal routes seem to be of good quality in terms of realism and precision. Since high resolution street networks were used for MM and to compute the optimal routes, origin and destination nodes are congruent and the LineStrings are smooth (see Figure 5.3). This high level of quality provides a solid basis for further analysis, especially for the calculation of distances and turn characteristics.

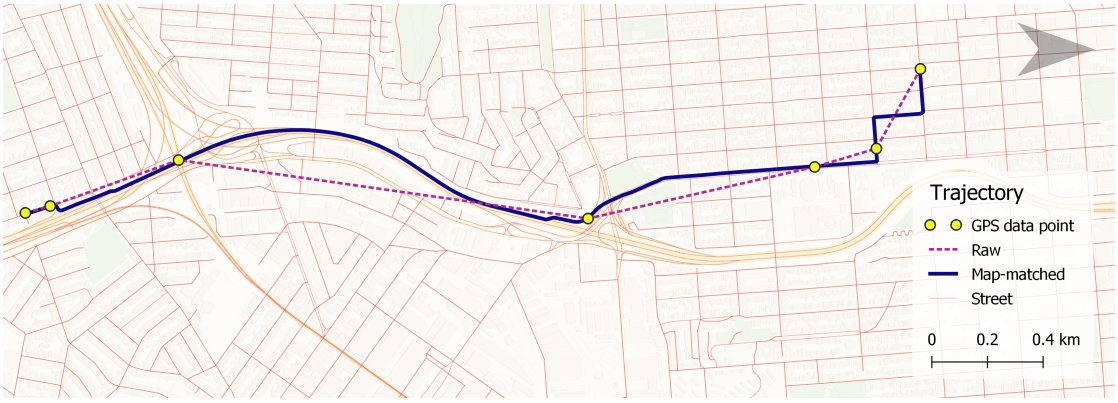


Figure 5.1: Exemplary map matching result from San Francisco showing the raw GPS data points, the raw trajectory and the map matched result on top of the street network.

Figure 5.2: Exemplary result of calculation of turns from an actual route in Vienna. Based on the angles between the segments in the LineString geometry, 8 turns were identified.[33] The marked curves were correctly classified as no turns.

[33] 5 right turns of which 3 are sharp turns and 3 left turns of which 2 are sharp turns.

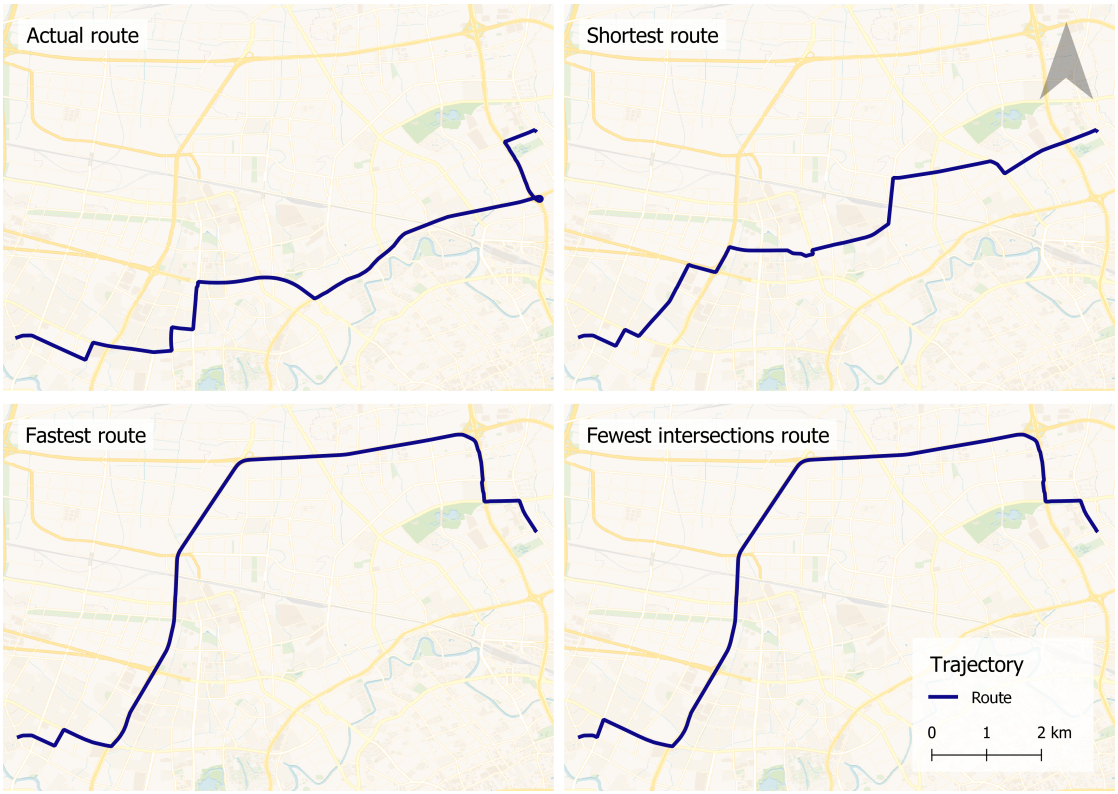


Figure 5.3: Example of actual and optimal routes for a trip in Shanghai where all four routes are different.

Figure 5.4 shows the data distribution of the three attributes which represent the criteria for the calculation of optimal routes: length, optimal duration, and number of intersections. A first overview shows that in general, the values in all three cities are similarly distributed with all skewness and kurtosis values indicating either a beta or a gamma distribution.[34] The differences between actual, shortest, fastest, and fewest intersections routes seem to be small as the curves are relatively congruent within each plot. Furthermore, the smoothness of the curves suggests that there are no extreme breaks or clusters present in the data.

When comparing the attributes between the three cities, the routes' lengths are distributed most similarly while the distributions of optimal durations and the numbers of intersections show differences in kurtosis and skewness. The Shanghai data stand out from the other two cities for both attributes as it shows the flattest distribution of optimal durations but the steepest distribution of numbers of intersections.

[34] The distribution of the data was assessed visually using skewness-kurtosis plots (Cullen and Frey 1999).

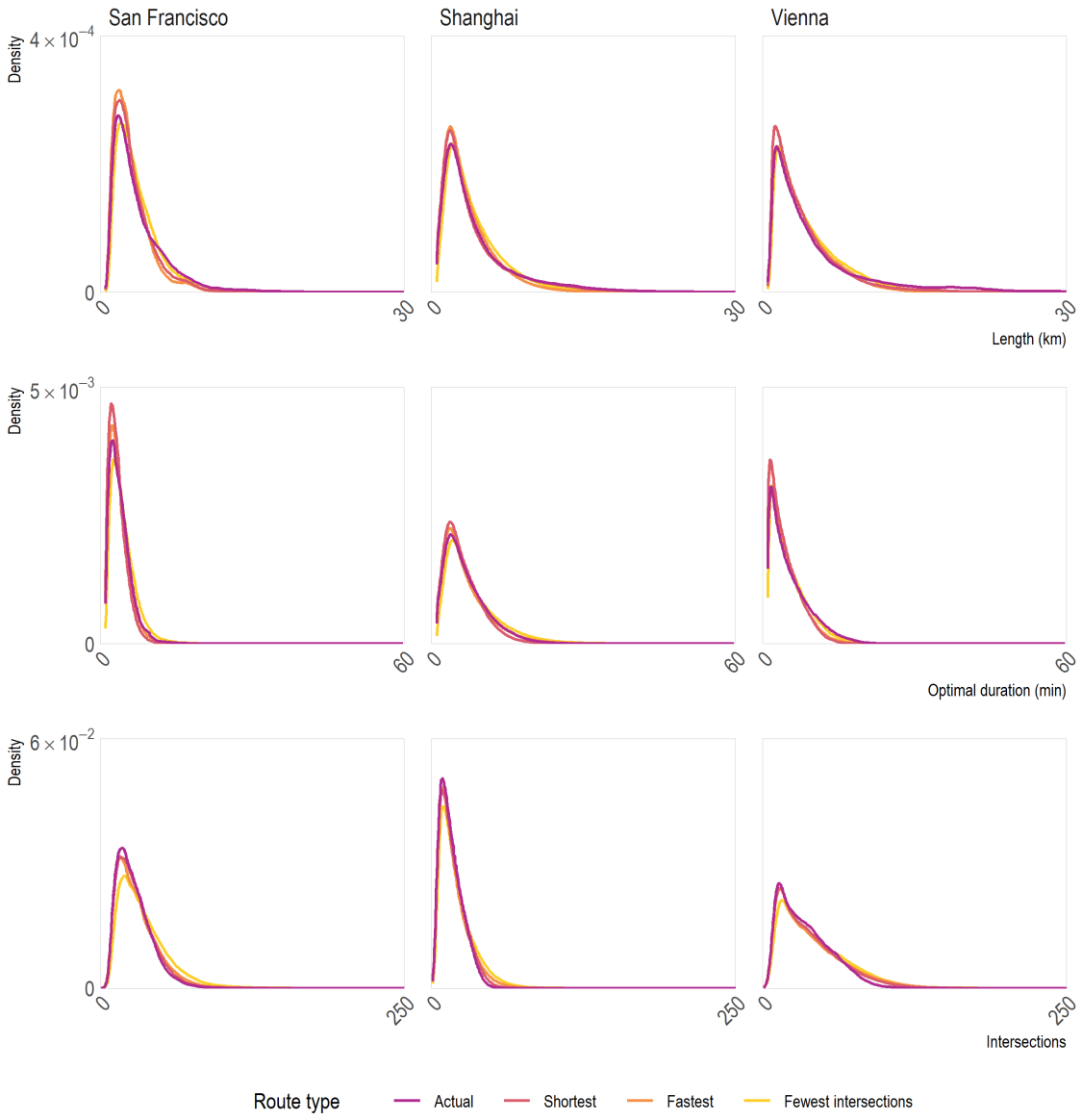


Figure 5.4: Density distributions of route characteristics.

Figure 5.5 shows the data represented in Figure 5.4 as boxplots[35]. Although some of the boxplots look very similar, most of the results are significantly different from each other.[36]

When comparing the results between the three cities, the first row of boxplots shows that most routes are shorter than 5 kilometres whereby routes in San Francisco are generally the shortest with routes in Shanghai and Vienna being 10 % and 17 %[37] longer, respectively. Surprisingly, a different and more extreme pattern is shown by the optimal duration boxplots in the second row suggesting that the median duration of routes in Shanghai is 69 % longer than in San Francisco and 45 % longer than in Vienna. Since the small difference in length between routes in San Francisco and Shanghai cannot explain this discrepancy, it is assumed that the duration values shown by the Shanghai boxplots are too high and therefore do not reflect driving behaviour. The most probable reason is the selection of generally lower speed limits for the calculation of edge travel times during network data preparation (see Table 4.1). However, this bias is not regarded a problem for further analysis as it does not affect the comparison of the different route types within each city. Regarding the difference between San Francisco and Vienna, the 17 % longer median route duration in Vienna corresponds to the length difference of 17 %. The bottom row of boxplots again reveals large differences between the three cities as the median number of intersections in Shanghai is 42 % lower than in San Francisco and 53 % lower than in Vienna. Since the route length is similar in all three cities, this means that the taxi drivers in Shanghai cross less intersections than their colleagues in San Francisco and Vienna. Although the intersection density in Shanghai is 4 to 5 times lower than in the other two cities (see Table 3.1), it cannot fully explain the results presented by the boxplots. The relationship between intersection density and the number of intersections a route travels is therefore further investigated and discussed in subsections 5.3.3 and 6.2.3.

A comparison of the results within each city reveals that the actual routes almost always present the largest median value. This indicates that optimal routes are not only optimal in terms of their specific characteristic but tend to optimise all three characteristics better than actual routes meaning that the fastest routes, for example, are not only faster than actual routes, but also tend to be shorter and containing less intersections. Furthermore,

[35] In all boxplots presented in this thesis, the black bar and its label represent the median and the hinges represent the 1st and 3rd quartile. The whiskers represent the soft outlier limits and extend at most $\pm 1.5 * IQR$, where IQR is the interquartile range. To avoid overloading the plots, outliers are not plotted.

[36] All data presented with a boxplot were examined for their central tendencies. For this purpose, a Mann-Whitney-U-Test (Wilcoxon 1945; Mann and Whitney 1947) was performed for each pair of data whereby the only result with a p-value larger than 0.0001 occurred when testing the route lengths of fastest routes in Shanghai and Vienna.

[37] These percentages represent the average of the individual percentage differences of the four route types.

the order of which route type shows the best median results for a certain criterion is almost identical in all three cities.
Regarding the ranges of values, Figure 5.5 reveals a general pattern that boxplots with high median values also show large Interquartile Ranges (IQRs) and long whiskers while boxplots with low median values show small IQRs and short whiskers. This supports the finding that the data do not contain extreme breaks or clusters.

Concluding this introductory section, it can be stated that the initial assessment of the dataset does not reveal any signs of incorrect data or data processing errors. The data are therefore supposed to be suitable for analysis and it is assumed that differences in the results are not caused by inconsistencies in the underlying data. The remainder of this chapter presents the results which are then discussed in Chapter 6.

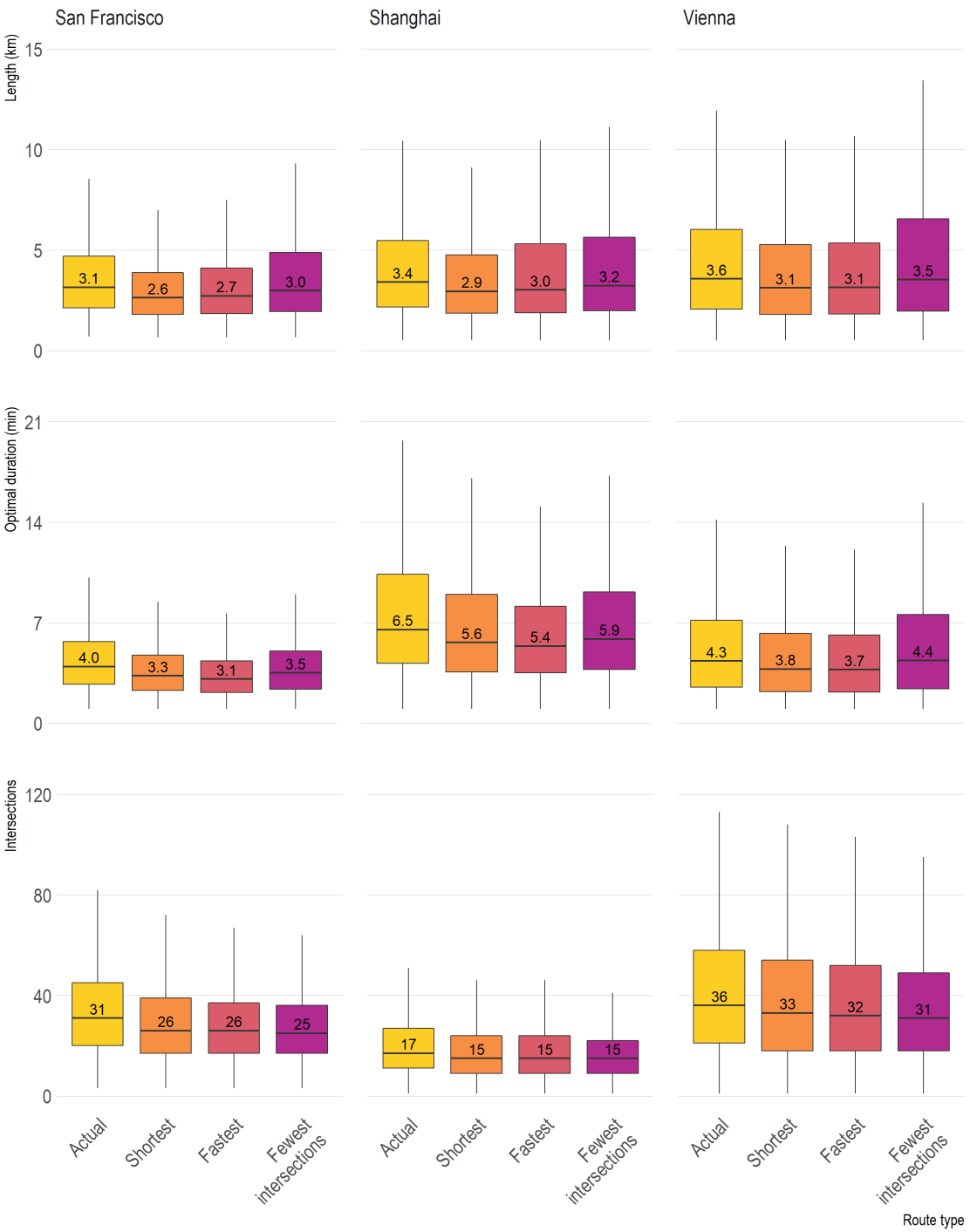


Figure 5.5: Boxplots of route characteristics. Note that the optimal duration does not reflect actual timescales since it was calculated based on speed limits and assumes perfect traffic conditions for all routes in order to make them comparable.

5.2 Similarity between actual and optimal routes

5.2.1 Percentage of shared length

This subsection presents the results from the analysis of route similarity between actual and optimal routes using the Percentage of Shared Length (PSL) (see Subsubsection 2.4.4.3 and Subsection 4.6.1). For a discussion of these results, see Subsection 6.1.1.

As an introduction, Figure 5.6 presents boxplots showing to what extent taxi drivers in each city follow the shortest, fastest, and fewest intersections routes.

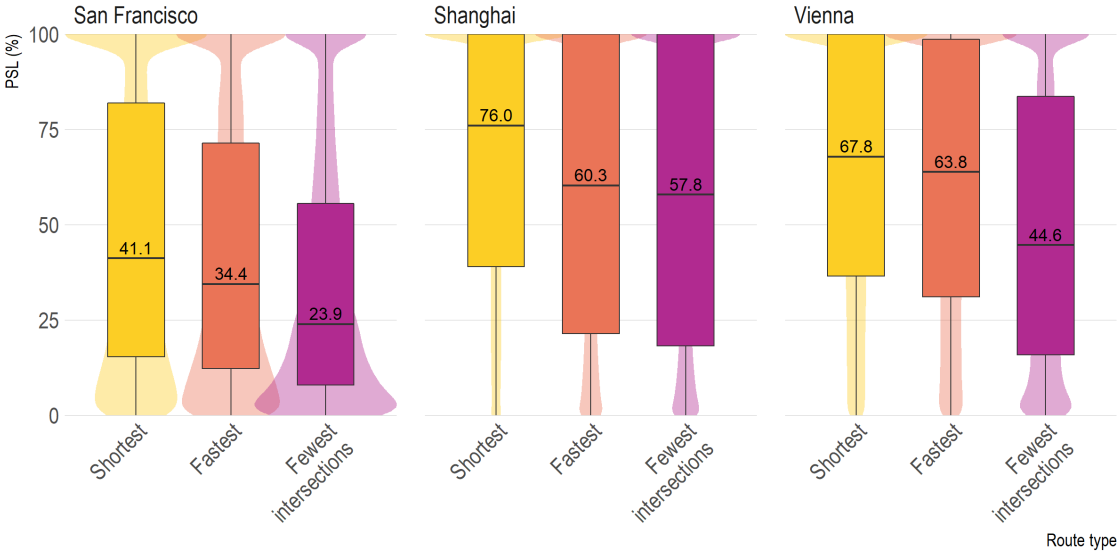


Figure 5.6: Boxplots of PSL. The area behind each boxplot qualitatively represents the data distribution.

The IQRs of the boxplots in Figure 5.6 contain about 50 % to 80 % of the data and the whiskers cover the entire data range which shows that the extent to which actual and optimal routes are similar varies greatly. The areas behind the boxplots show that San Francisco differs from the other two cities in terms of the results' distributions as it shows that the results contain a relatively large number of high and low values. In contrast, the results in Shanghai and Vienna do not contain remarkably many low values but only a large proportion of high values. This difference in data distribution also explains the generally lower median values in San Francisco which reveal that the median route shares 41.1 % of its length with the shortest, 34.4 % with the fastest, and 23.9 % with the fewest intersections alternative. The routes from Shanghai present considerably larger PSL values indicating that the median route is 76.0 % congruent with the shortest, 60.3 % congruent with the fastest, and 57.8 % identical to the fewest intersections routes.

The routes from Vienna share 67.8 % of their length with the shortest, 63.8 % with the fastest, and 44.6 % with the fewest intersections route. It is apparent that all routes are most similar to shortest routes and least similar to fewest intersections routes. It is also worth mentioning that the relative difference between the route types is not the same in all cities: in San Francisco, all three boxplots are clearly different whereas in Shanghai, the results of the fastest and fewest intersections routes are similar and in Vienna, it is the boxplots of shortest and fastest routes showing similarities.
A more detailed overview of the PSL results is presented in Table 5.1 which shows the percentage of actual routes in given PSL intervals. It shows that only 18.1 % of drivers in San Francisco follow exactly the shortest route and that this proportion is even lower in terms of fastest and fewest intersection routes. The results from San Francisco further show that more than half of the routes overlap less than 50 % with the shortest or fastest alternatives and less than 25 % with the fewest intersections alternative. The high values in the 0–25 % interval indicate that a large proportion of drivers choose a route that differs greatly from the optimal route. In contrast to the taxi drivers in San Francisco, the drivers in Shanghai seem to follow optimal routes more consistently as over a third of all routes are identical with the shortest, fastest, or fewest intersections alternative. However, even in Shanghai, between 15.9 % and 30.4 % of the routes share less than a quarter of their distance with an optimal route. It is noticeable that the proportion of routes with a PSL of 100 % is very similar for all three route types which is not the case in the 0–25 % interval. The bottom third of Table 5.1 shows that in Vienna, about 25 % of the routes are identical to the shortest or fastest route but only 16.8 % follow the fewest intersections route. More than 60 % of the routes correspond to at least 50 % with the shortest or fastest alternative while only 46.1 % of the routes share more than half of their distance with the fewest intersections route. Similar to Shanghai, between 16.8 % and 34.1 % of the routes in Vienna hardly correspond to any of the optimal routes.

Table 5.1: Percentage of optimal routes in different intervals of PSL between actual trips and optimal routes.

PSL (%)	**Percentage of routes in interval (%)**					
	Shortest		**Fastest**		**Fewest intersections**	
San Francisco	*	**	*	**	*	**
100	18.13	*18.13*	13.46	*13.46*	9.58	*9.58*
75 – 100	10.11	*28.24*	9.74	*23.20*	6.82	*16.40*
50 – 75	14.84	*43.08*	14.38	*37.58*	11.76	*28.16*
25 – 50	21.47	*64.55*	21.38	*58.95*	20.52	*48.68*
0 – 25	35.45	*100.00*	41.05	*100.00*	51.32	*100.00*
Shanghai	*	**	*	**	*	**
100	38.82	*38.82*	33.93	*33.93*	34.01	*34.01*
75 – 100	11.83	*50.65*	8.15	*42.08*	7.57	*41.58*
50 – 75	16.92	*67.56*	13.97	*56.05*	12.62	*54.19*
25 – 50	16.45	*84.01*	16.20	*72.25*	15.41	*69.60*
0 – 25	15.99	*100.00*	27.75	*100.00*	30.40	*100.00*
Vienna	*	**	*	**	*	**
100	25.57	*25.57*	24.24	*24.24*	16.88	*16.88*
75 – 100	18.39	*43.96*	17.05	*41.28*	13.34	*30.22*
50 – 75	20.66	*64.62*	19.30	*60.59*	15.92	*46.13*
25 – 50	18.51	*83.13*	18.94	*79.53*	19.73	*65.86*
0 – 25	16.87	*100.00*	20.47	*100.00*	34.14	*100.00*

* regular columns show percentages of optimal routes per PSL interval

** cursive columns show cumulative percentages

Figure 5.7 visualises the relationship between PSL and Origin-Destination Distance (ODD). When looking at the first row showing the results of the San Francisco routes, it is noticeable that the trends in all three plots are very similar. This indicates that the drivers' preferences as shown in Figure 5.6, namely that their routes most closely resemble the shortest routes, are not dependent on the length of the trip. The San Francisco boxplots further all show the same pattern that PSL values decrease with increasing ODD as long as the ODD is less than 5 kilometres, but then increase again when trips become longer and have ODDs between 5 and 8 kilometres. If the routes then become even longer, the PSL decreases again when ODD increases. However, it needs to be mentioned that the data in this ODD range is sparse as only 2.7 % of the routes in the dataset have an ODD longer than 8 kilometres. The second row of boxplots shows that in Shanghai, the PSL generally decreases as the routes become longer. However, this trend is much less strong for the shortest routes which suggests that the length of the chosen route is particularly important to taxi drivers when origin and destination are far apart. The Shanghai plots further show that the PSL of fastest and fewest intersections routes stabilises at a low level when ODDs are larger than about 7 kilometres. The bottom row of boxplots reveals that the trends in Vienna are very similar to those in Shanghai. However, there is one striking difference, namely that the similarity between actual and fewest intersections routes increases again with ODD when the actual ODDs are longer than 12 kilometres. This trend is comparable to the one revealed by the results from San Francisco but is also based on very few routes as less than 0.04 % of the routes in Vienna have an ODD of more than 12 kilometres.

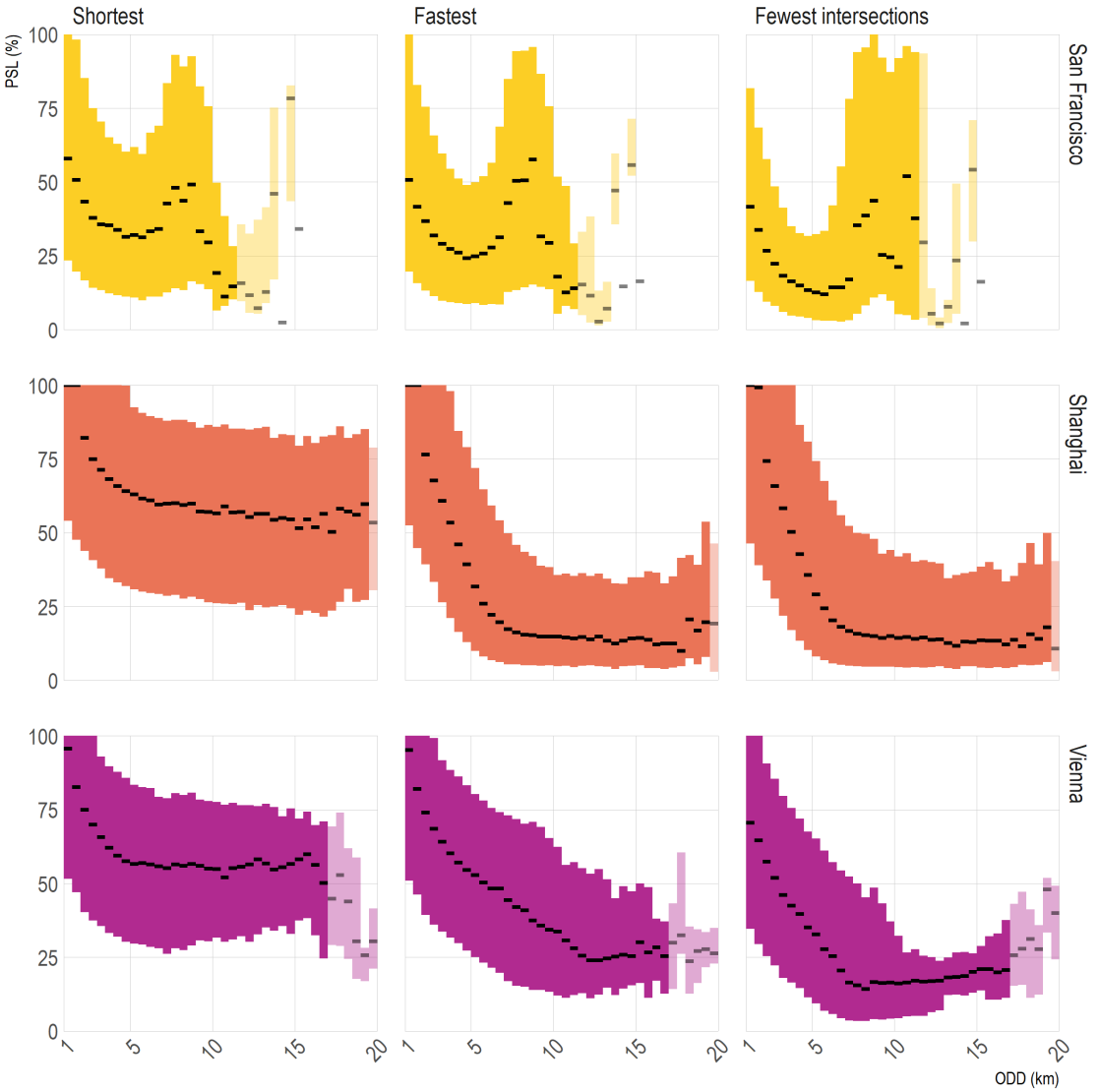


Figure 5.7: Boxplot series of PSL over ODD. The data was grouped into ODD intervals of 500 metres whereby each interval's median and IQR are indicated by a black bar and a coloured box, respectively. Semi-transparent colors indicate that the respective interval contains less than 100 routes.

Figure 5.8 presents the strength and significance of the trends mentioned above. It confirms their significance, in particular that the increase and renewed decrease of PSL, which seems to take place between ODDs of 6 and 12 kilometres in San Francisco, is significant. Furthermore, the mentioned positive correlation between the PSL of fewest intersections routes and the ODD in Vienna for very long routes is confirmed.

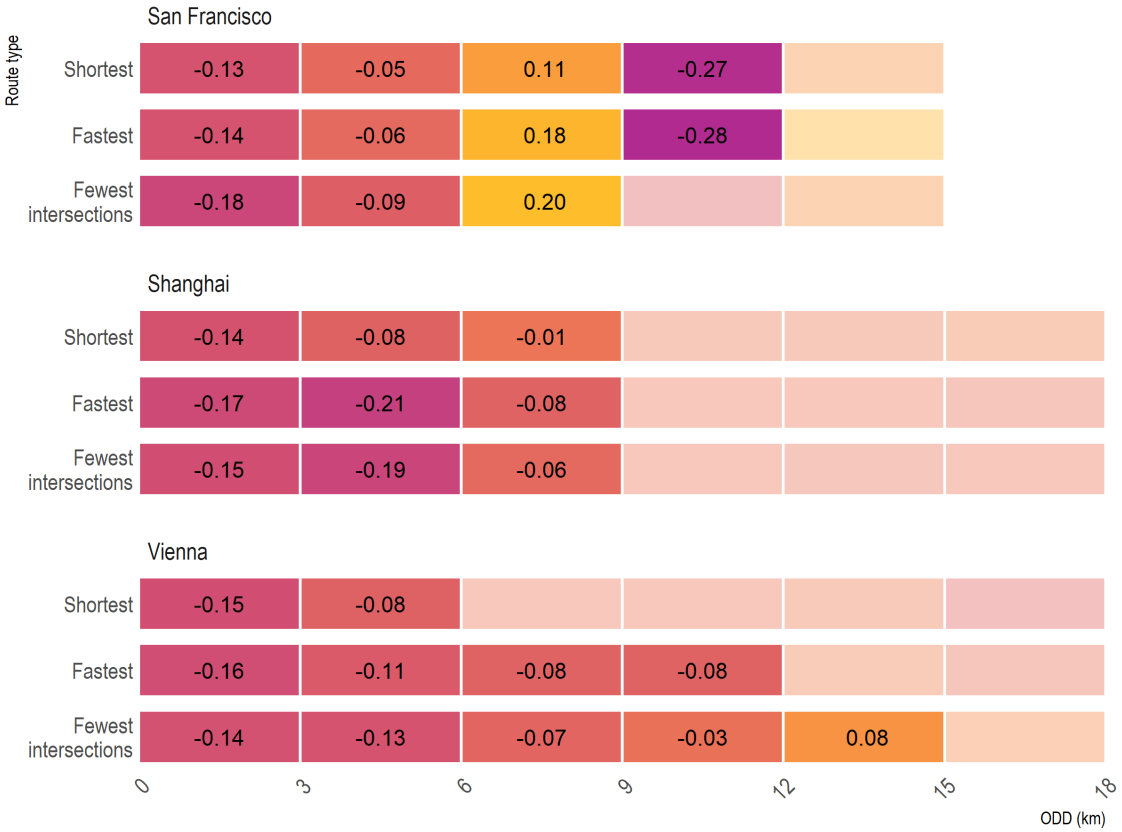


Figure 5.8: Correlation of PSL with ODD. The values represent the Spearman rank correlation coefficient ρ (Spearman 1904) which was used to calculate the correlation within each given ODD interval. According to Cohen (1992), $\rho = \pm 0.1$ corresponds to a weak effect, while $\rho = \pm 0.3$ implies a medium effect, and $\rho = \pm 0.5$ represents a strong effect. Dark values indicate negative and bright values indicate positive correlation. Semi-transparent tiles indicate that there is no significant correlation (p-value > 0.0001).

Figure 5.9 visualises how the PSL differs during the day. The first two rows of tileplots, representing the results from San Francisco and Shanghai, show no clear pattern except that the PSL values differ in almost all time intervals. The Vienna results, however, show clearly that the PSL values of routes conducted during the night are significantly different from the PSL of routes conducted during the day.

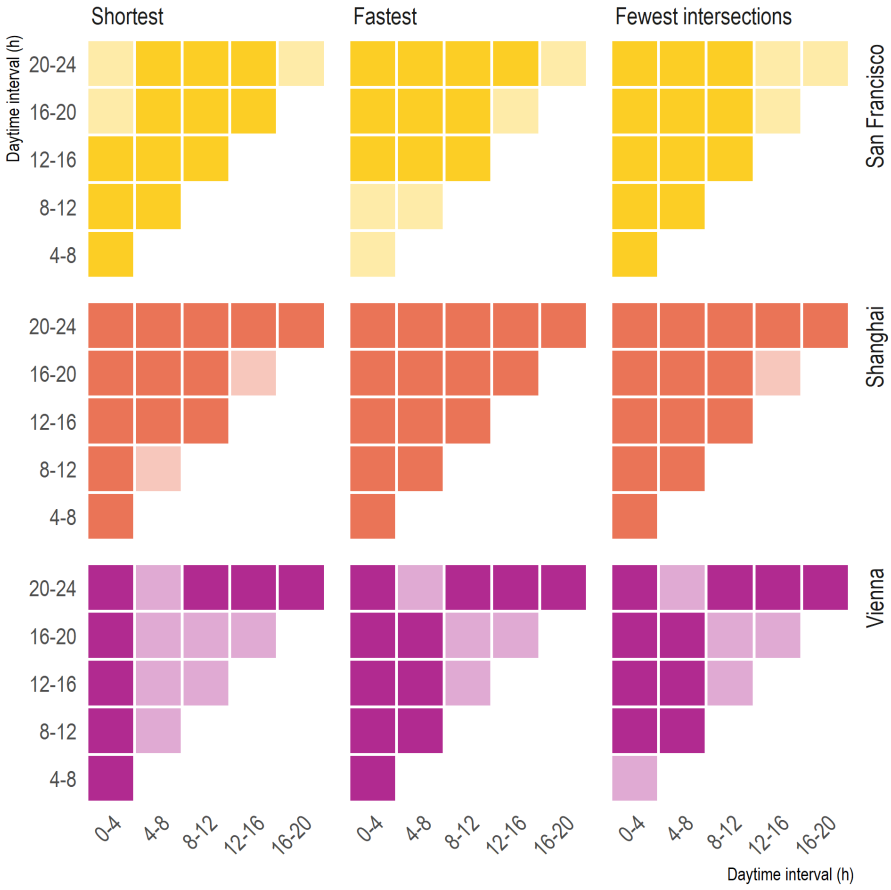


Figure 5.9: Differences in PSL between different time periods. The differences were assessed with pairwise Mann-Whitney-U-Tests (Wilcoxon 1945; Mann and Whitney 1947) whereby each square in the plot represents a result. Non-transparent tiles indicate significance and semi-transparent tiles indicate that there is no significant difference (p-value > 0.0001).

Similar to Figure 5.9, Figure 5.10 shows which days of the week differ in terms of the PSL between actual and optimal routes. The results from San Francisco suggest that there is a difference between weekends and weekdays as the PSL values from routes conducted on a Saturday or Sunday are not significantly different from each other but different from all the other weekdays. A similar pattern is revealed by the second row of tileplots indicating that in Shanghai, routes driven on a Thursday or Friday differ from routes conducted on another day of the week. The plots in the bottom row show that the PSL of routes in Vienna is not depending on the weekday as none of the days displays significant difference.
The patterns revealed by figures 5.9 and 5.10 were further investigated and the results are presented in figures 5.11, 5.12, and 5.13.

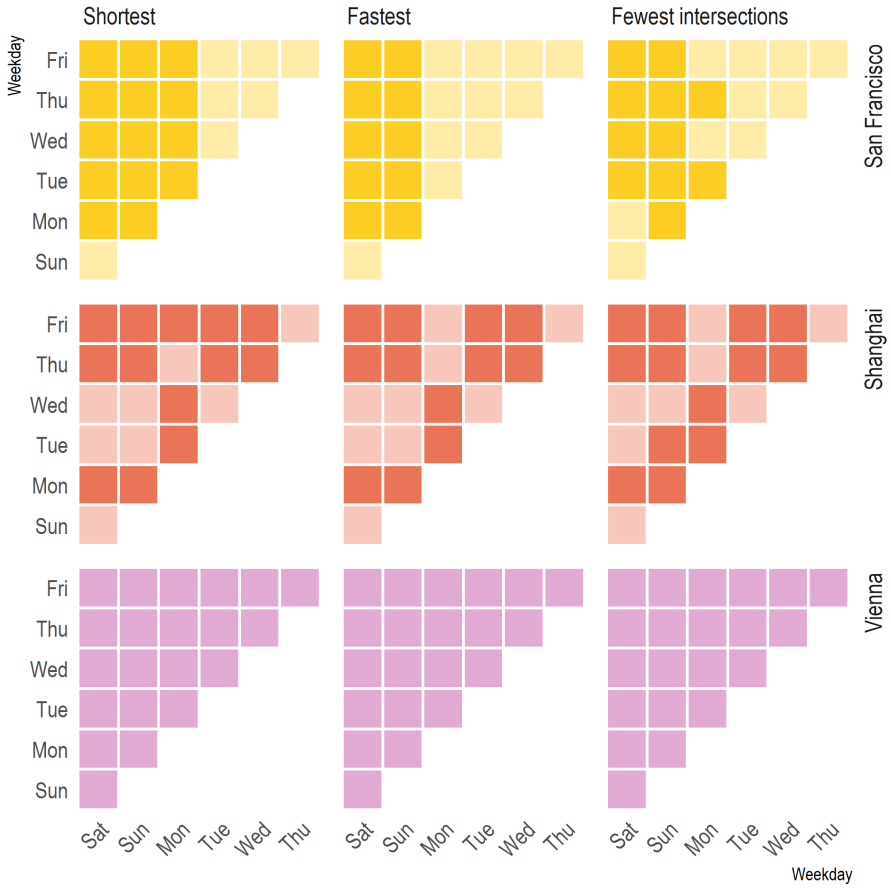


Figure 5.10: Differences in PSL between different weekdays. The differences were assessed with pairwise Mann-Whitney-U-Tests (Wilcoxon 1945; Mann and Whitney 1947) whereby each square in the plot represents a result. Non-transparent tiles indicate significance and semi-transparent tiles indicate that there is no significant difference (p-value > 0.0001).

Figure 5.11 shows a comparison of the PSL from routes in San Francisco for weekdays and weekends. The three plots suggest that the PSL of all three route types tends to be lower during the weekend which indicates that taxi drivers choose less optimal routes during the weekend. The differences are largest regarding the PSL between actual and shortest routes but relatively small with regard to fastest and fewest intersections routes.

Figure 5.11: Comparison of PSL over ODD on weekends and weekdays in San Francisco. The data was grouped into ODD intervals of 500 metres whereby each interval's median is represented with a dot. Only intervals containing more than 100 routes are plotted. The lines between the dots do not represent data but are for visualisation only.

Figure 5.12 compares the relationship between PSL and ODD in Shanghai on Thursdays and Fridays with the other weekdays' results and reveals that although a significant difference was found, both lines show an almost identical course. It is therefore assumed that the PSL of actual routes in Shanghai does not vary considerably on different days.

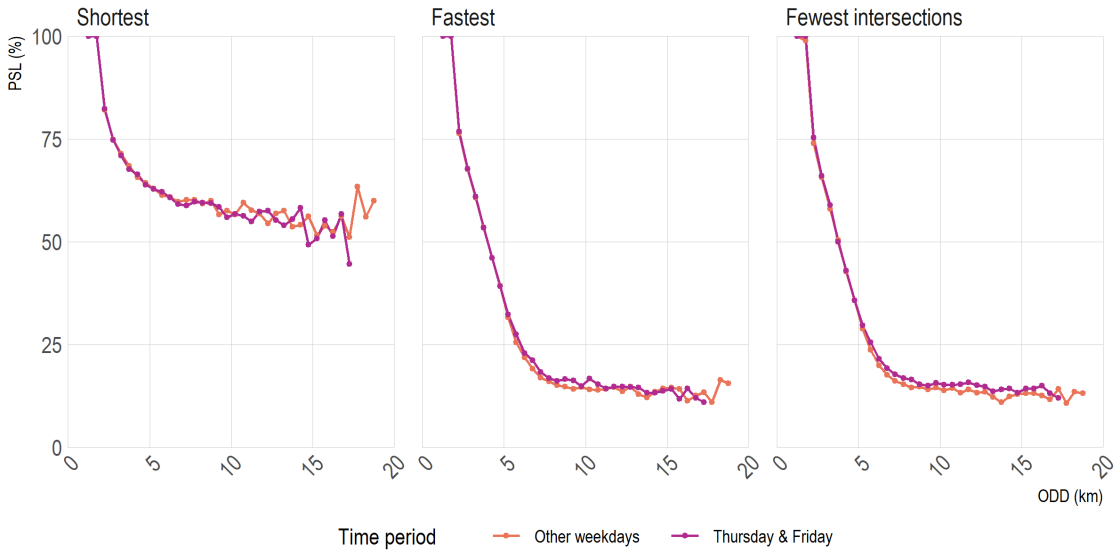


Figure 5.12: Comparison of PSL over ODD on different weekdays in Shanghai. The data was grouped into ODD intervals of 500 metres whereby each interval's median is represented with a dot. Only intervals containing more than 100 routes are plotted. The lines between the dots do not represent data but are for visualisation only.

Figure 5.13 shows the differences between day and night in terms of PSL between actual and optimal routes in Vienna. The trends in all three plots are almost identical when ODDs are smaller than about 6 kilometres but differ when ODDs are longer. The results suggest that taxi drivers chose routes which are less congruent with the alternatives during the night. However, this is only true for the fastest and fewest intersections routes with ODDs between 6 and 12 kilometres.

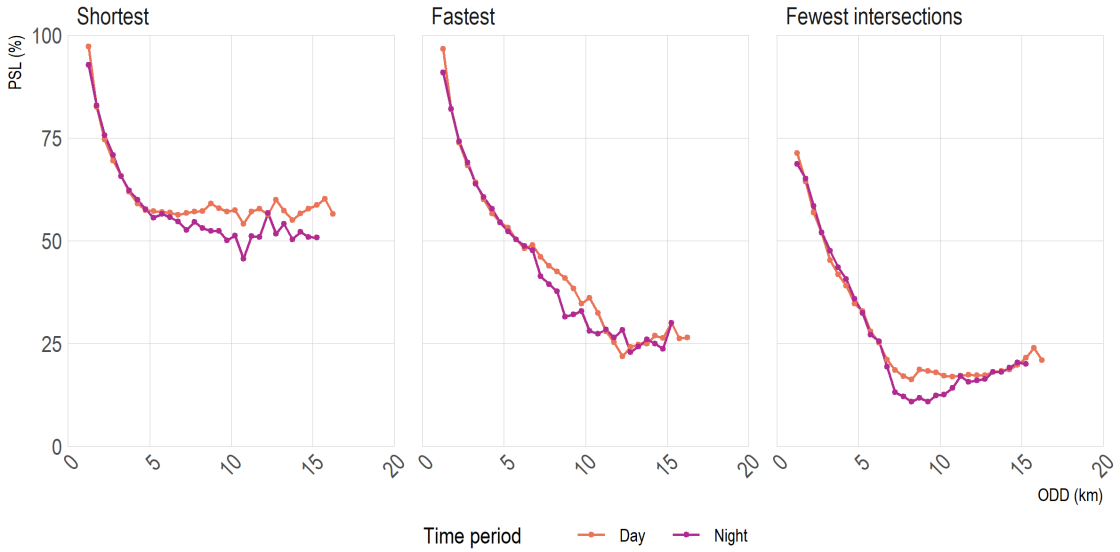


Figure 5.13: Comparison of PSL over ODD during day (4 am – 8 pm) and night (8 pm – 4 am) in Vienna. The data was grouped into ODD intervals of 500 metres whereby each interval's median is represented with a dot. Only intervals containing more than 100 routes are plotted. The lines between the dots do not represent data but are for visualisation only.

The results presented in this subsection can be summarised as follows:

General results

- The extent to which taxi drivers' routes are identical with optimal routes differ between the three cities but also between individual trips.
- In general, routes from Shanghai are most similar to optimal routes while the routes from San Francisco are least similar to optimal routes.
- In all three cities, taxi drivers' routes are most similar with the shortest alternative.

Results regarding a specific city

- In San Francisco, taxi drivers prefer to optimise their routes for distance no matter the length of a trip, but it is surprising that they seem to choose more optimal routes when the ODD is around 7 kilometres than when the ODD is 2 kilometres.
- Taxi drivers in Shanghai and Vienna choose routes which are most similar to the shortest alternative, especially when trips are long.
- During the night, taxi drivers in Vienna choose routes which are less similar to the optimal routes if the ODD is longer than 6 kilometres.
- In San Francisco and Shanghai, there seems to be no substantial difference between day and night regarding route similarity between actual and optimal routes.
- Routes in Shanghai and Vienna do not differ on different days regarding their similarity to optimal routes.

For a discussion of these results, see Subsection 6.1.1.

5.2.2 Fréchet distance

The Fréchet Distance (FD) was computed as an alternative measure of route similarity (see Subsubsection 2.4.4.3 and Subsection 4.6.1). However, most of the analysis is based on the Percentage of Shared Length (PSL) which is why the results of the FD are only used to verify the trends identified by the PSL. Furthermore, the FD is not an intuitive measure of similarity which is why this subsection focuses on relative differences and the relationship to the PSL but does not compare absolute values. For a discussion of the results presented here, see Subsection 6.1.1.

Figure 5.14 shows that the boxplots are generally similar and that the IQRs are relatively small in comparison to the range of values which indicates that the extent to which actual and optimal routes are alike is similar between and within the cities. The areas behind the boxplots reveal that most values are low but not 0.

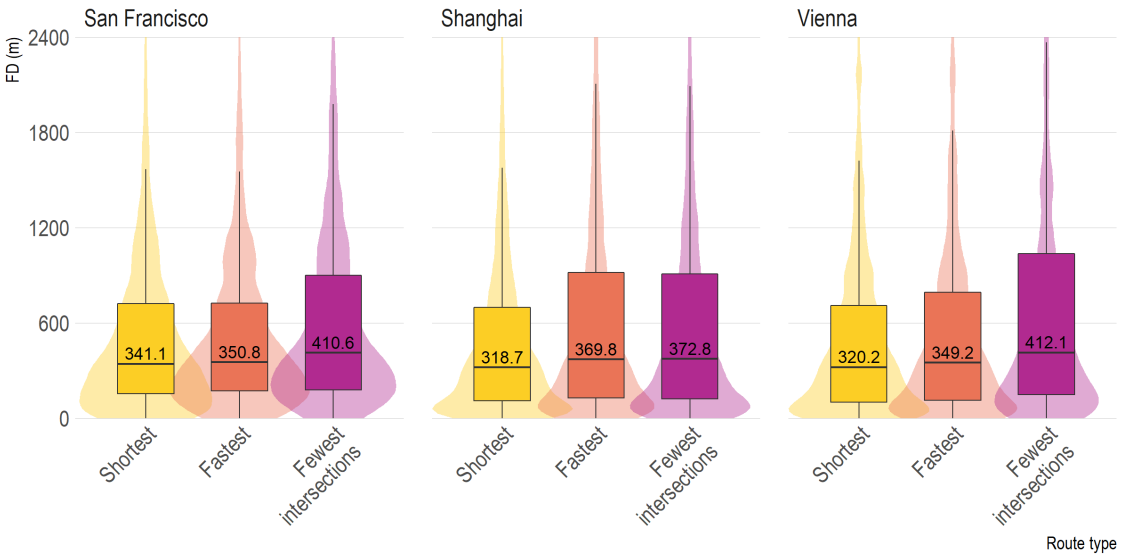


Figure 5.14: Boxplots of FD. The area behind each boxplot qualitatively represents the data distribution.

Figure 5.15 visualises the relationship between FD and Origin-Destination Distance (ODD). The first row, showing the results from San Francisco, presents three similar plots which all show that the FD increases with increasing ODD. However, this rise in FD decreases for fastest and fewest intersections routes when the routes have ODDs of more than 6–8 kilometres. These trends suggest that in San Francisco, the similarity between actual and optimal routes generally decreases with increasing trip length with taxi drivers tending to choose a route which is similar to the fastest or fewest intersections route when the ODD of their trip is longer than about 10 kilometres. The second row of plots, presenting the results from Shanghai, shows that in general, the similarity between actual and optimal routes in Shanghai also decreases with increasing ODD. However, the decrease is less strong for shortest routes than for fastest and fewest intersections routes. This shows that taxi drivers in Shanghai tend to optimise their routes by distance, especially if the route is long. The results from Vienna show the same trend as those from Shanghai with the difference that the increase in FD for fewest intersections routes stops when the routes' ODDs are longer than about 12 kilometres.

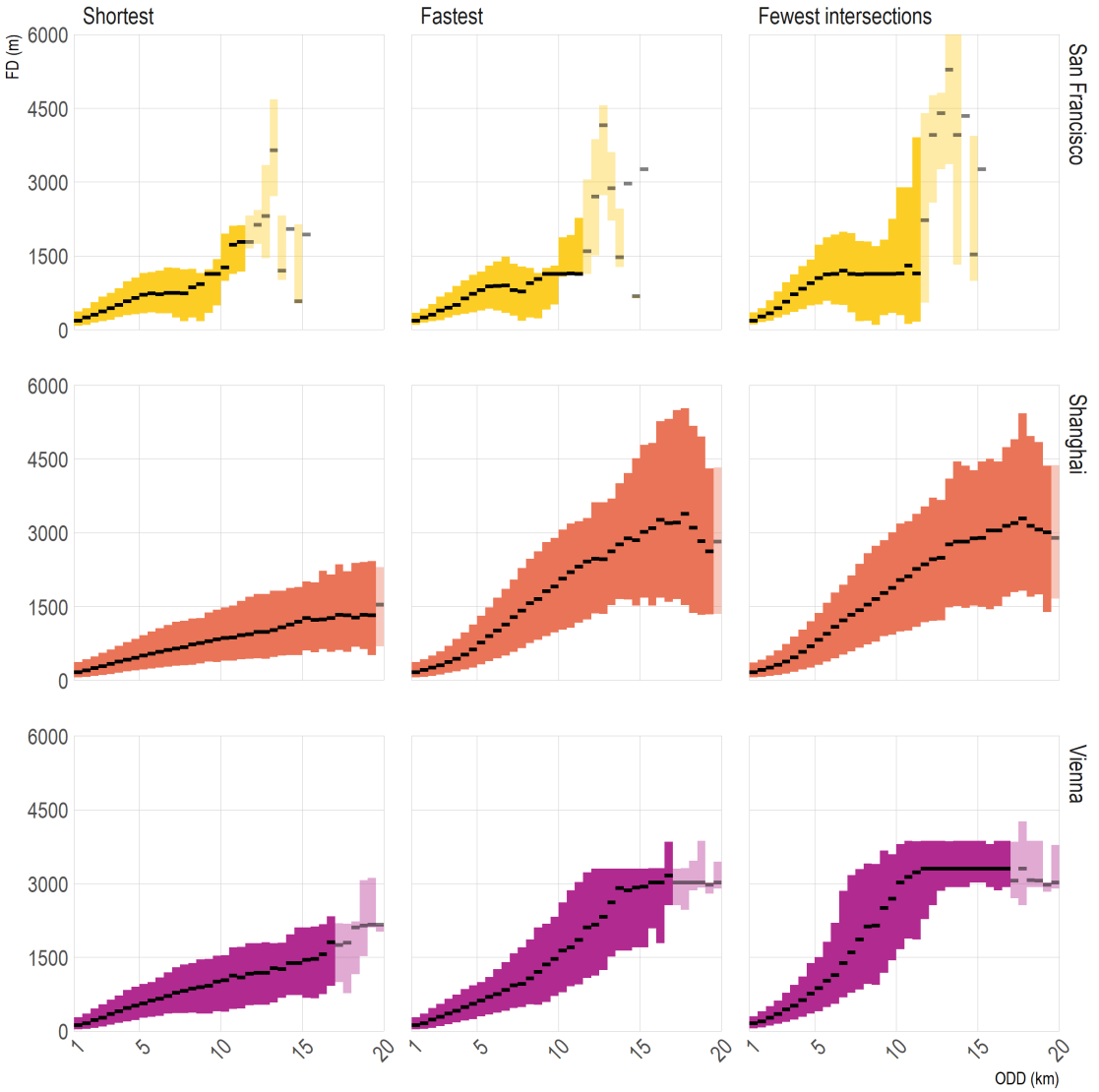


Figure 5.15: Boxplot series of FD over ODD. The data was grouped into ODD intervals of 500 metres whereby each interval's median and IQR are indicated by a black bar and a coloured box, respectively. Semi-transparent colors indicate that the respective interval contains less than 100 routes.

Overall, the results presented by Figure 5.15 show the same trends as the results from the PSL presented in Figure 5.7. Nevertheless, the relationship between FD and PSL is shown again in Figure 5.15 which shows opposite trends in all plots as a high similarity is represented by high PSL but low FD values. The relationship revealed by Figure 5.16 was verified using the Spearman rank correlation coefficient ρ (Spearman 1904) which revealed strong significant negative correlation between the two similarity measures in

all ODD intervals.[38]

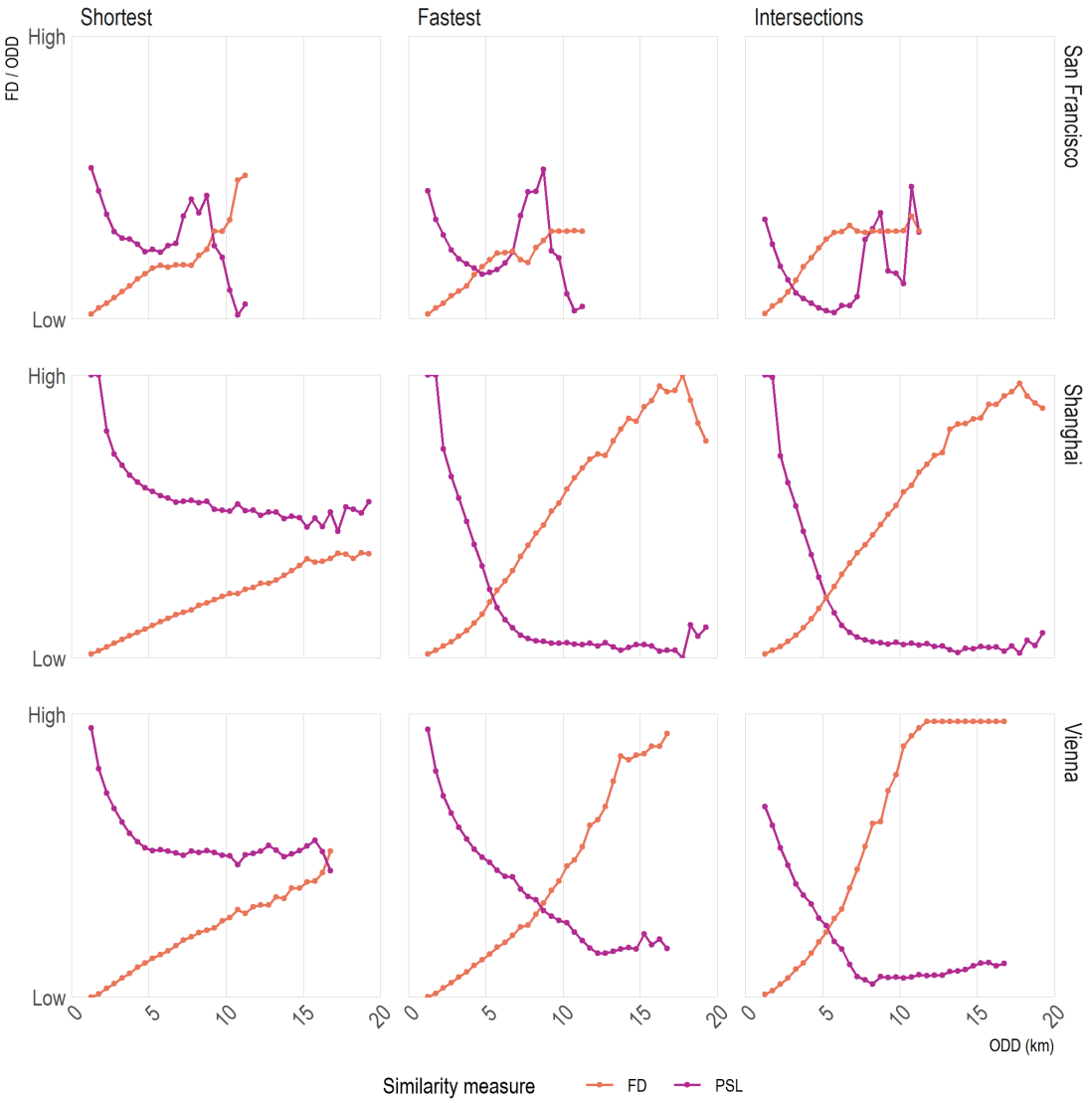


Figure 5.16: Comparison of FD and PSL over ODD. The data was grouped into ODD intervals of 500 metres whereby each interval's median is represented with a dot. The values were rescaled to an interval of [0,1] for visualisation. Only intervals containing more than 100 routes are plotted. The lines between the dots do not represent data but are for visualisation only.

[38] To compute correlation, the data was grouped into ODD intervals of 3 kilometres and then, correlation was calculated within each interval. All values were significant (p-value$\leq$ 0.0001) and between -0.45 and -0.88 which corresponds to strong correlation according to Cohen (1992).

The results presented in this subsection can be summarised as follows:

General results

- The extent to which actual and optimal routes are alike is similar between and within the cities.
- The relationship between FD and ODD supports the trends revealed by the analysis of PSL (see Subsection 5.2.1).

The results presented in this subsection are not discussed separately but are included in the discussion of the results from the analysis of PSL (see Subsection 6.1.1).

5.2.3 Percentage difference of length, duration, and number of intersections

This section presents the results of the comparison between actual and optimal routes in terms of length, duration, and number of intersections using the Percentage of Length Difference (PLD), Percentage of Time Difference (PTD), and Percentage of Intersections Difference (PID) (see Subsubsection 2.4.4.3 and Subsection 4.6.1). For a discussion of the results presented here, see Subsection 6.1.2.
In all plots presented in this subsection, positive values indicate that actual routes perform worse meaning that they are longer, slower, or include more intersections than the respective alternatives. Furthermore, it needs to be pointed out that the PTD is not based on actual but on optimal durations[39]. Thus, differences revealed by the PTD are not caused by the way travel times were determined but by taxi drivers choosing routes which diverge from the fastest route.

[39] The optimal duration is the time a vehicle would need to cover the given route if it could drive with the maximum speed allowed and without slowing down on intersections, traffic lights, or crossings (see Section 4.6).

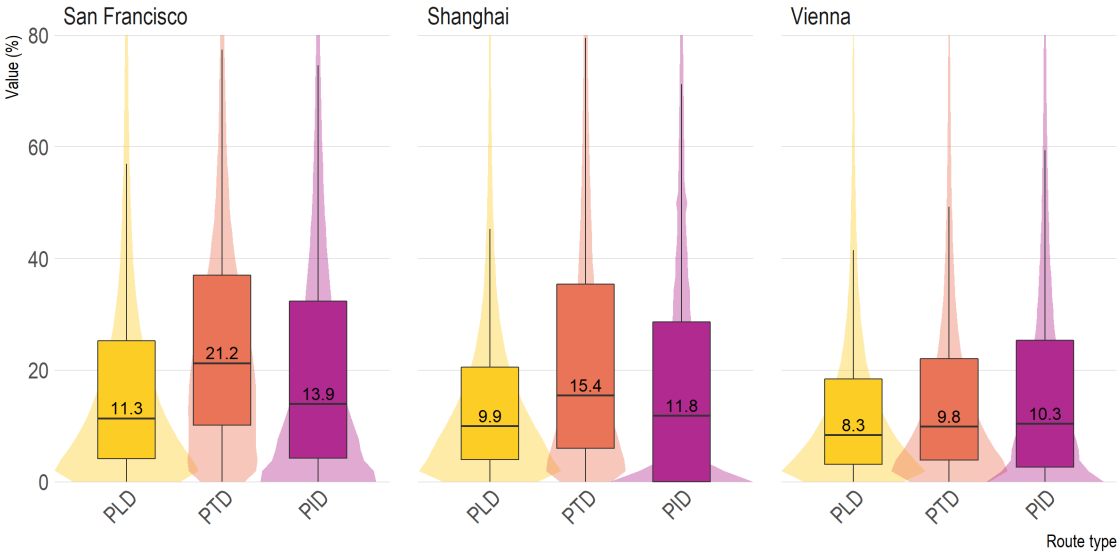


Figure 5.17: Boxplots of PLD, PTD, and PID. The area behind each boxplot qualitatively represents the data distribution.

Figure 5.17 visualises how much less efficient actual routes are compared to optimal routes. It shows that in general, the medians are relatively low in relation to the whiskers which means that there are mostly low but also some high values. This is also illustrated by the areas behind the boxplots representing data distributions. Figure 5.17 further shows that the PLD presents the lowest results in all three cities and reveals that San Francisco presents the highest values while Vienna shows the lowest results. San Francisco also presents the largest differences between the different route characteristics with the median PTD being almost double as high as the median PLD. The first three boxplots show that in San Francisco, a median route is 11.2 % longer than the shortest route and 21.1 % slower than the fastest route, including 13.9 % more intersections than the fewest intersections alternative. Shanghai shows the same pattern of actual routes deviating least in length and most in duration from the optimal alternative. Compared to the results from San Francisco, the median values of PLD, PTD, and PID are lower and more similar, showing that the median route in Shanghai is only 9.9 % longer than the shortest and 15.4 % slower than the fastest route, crossing 11.8 % more intersections than the fewest intersections alternative. As mentioned, Vienna presents the lowest and most similar median values with the median route being 8.3 % longer than the shortest and 9.8 % slower than the fastest route, including 10.3 % more intersections than the fewest intersections route.

Table 5.2 presents the percentage of optimal routes in different intervals of PLD, PTD, and PID. It shows that 28.7 % of the actual routes in San Francisco are less than 5

% longer than the shortest alternative and almost half of the routes are at most 10 % longer than the shortest route. In terms of intersections, the results are similar with 27.9 % of the taxi drivers' routes containing less than 5 % more intersections than the optimal alternative. However, in terms of route duration, only 13.6 % of the actual routes take less than 5 % longer than the fastest alternative and more than half of the routes chosen by taxi drivers in San Francisco take over 20 % longer than the fastest route. The routes chosen by taxi drivers in Shanghai are similar to the ones in San Francisco with 29.8 % of the drivers choosing a route which is less than 5 % longer and over half of them taking a route which is at most 10 % longer than the most direct route. The drivers in Shanghai perform better in terms of number of intersections with 32.8 % of them choosing a route which includes less than 5 % more intersections than the fewest intersections alternative. The largest difference between San Francisco and Shanghai is presented by the PTD which shows that in Shanghai, a much larger proportion of the actual routes takes less than 5 % longer than the fastest route. As already shown in Figure 5.17, Vienna presents the most homogeneous results with the proportion of actual routes deviating by less than 5 % from the optimal route being approximately one third for all three route characteristics. The aspect setting Vienna apart from the other two cities is that a relatively high proportion of actual routes is only slightly longer than the fastest route.

Table 5.2: Percentage of optimal routes in different intervals of PLD, PTD, and PID.

Interval (%)	Percentage of routes in interval (%)					
	PLD		PTD		PID	
San Francisco	*	**	*	**	*	**
0 – 5	28.79	*28.79*	13.62	*13.62*	27.92	*27.92*
5 – 10	17.62	*46.41*	11.29	*24.91*	14.27	*42.19*
10 – 20	21.74	*68.15*	22.71	*47.62*	18.78	*60.97*
20 – 40	17.13	*85.29*	30.25	*77.87*	20.17	*81.14*
40 – ...	14.71	*100.00*	22.13	*100.00*	18.86	*100.00*
Shanghai	*	**	*	**	*	**
0 – 5	29.86	*29.86*	21.64	*21.64*	32.83	*32.83*
5 – 10	20.36	*50.23*	15.53	*37.17*	13.76	*46.59*
10 – 20	24.04	*74.27*	20.96	*58.13*	19.58	*66.16*
20 – 40	15.69	*89.96*	20.15	*78.28*	17.28	*83.44*
40 – ...	10.04	*100.00*	21.72	*100.00*	16.56	*100.00*
Vienna	*	**	*	**	*	**
0 – 5	35.56	*35.56*	30.94	*30.94*	33.48	*33.48*
5 – 10	20.27	*55.83*	19.66	*50.60*	16.09	*49.57*
10 – 20	21.42	*77.25*	21.81	*72.41*	19.19	*68.77*
20 – 40	13.82	*91.07*	15.88	*88.29*	17.41	*86.17*
40 – ...	8.93	*100.00*	11.71	*100.00*	13.83	*100.00*

* regular columns show percentages of optimal routes per PLD, PTD, or PID interval
** cursive columns show cumulative percentages

Figure 5.18 visualises how the PLD, PTD, or PID are related to the Origin-Destination Distance (ODD). It reveals a common trend that the PLD does barely change when ODDs increase which means that actual routes in San Francisco are always about 13 % longer than the shortest alternative while the difference is about 8 % in Shanghai and Vienna. The second column of boxplots shows very different trends of PTD: in San Francisco, the PTD is relatively stable at about 20 % when ODDs are lower than 7 kilometres but varies when ODDs become longer. Although the data are sparse, as

only 4.3 % of the routes in San Francisco have an ODD above 7 kilometres, the boxplot indicates that the PTD drops and increases again when ODDs are between 7 and 10 kilometres which means that a route with an ODD of 9 kilometres tends to diverge less from the fastest route, in terms of duration, than a route which is only a few kilometres long. In Shanghai, the PTD shows a strong and steady increase suggesting that the longer an actual route is, the slower it is in relation to the fastest route whereby routes with an ODD above 10 kilometres, which accounts for 2.6 % of the routes, are likely to be over 50 % slower than the fastest route. Vienna shows the same general trend of steadily increasing PTD values but the increase is much slower so that even very long routes should not take more than 25 % longer than the fastest alternative. The PID boxplots reveal similar patterns as the PTD boxplots. However, the trends are more extreme in San Francisco and Vienna and it is worth mentioning that the mentioned de- and increase of PTD in San Francisco can also be seen in the PID results where it is even stronger.

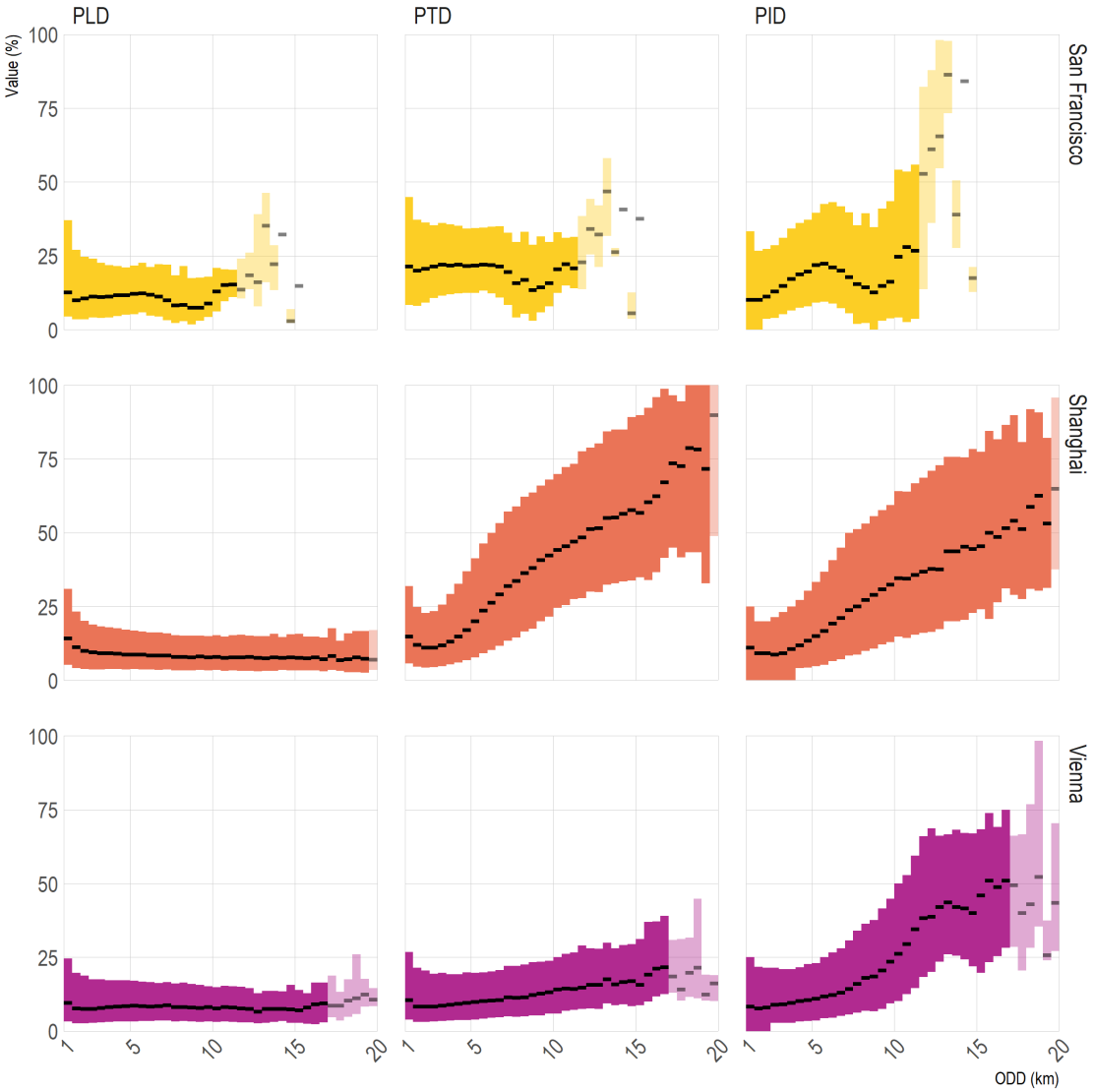


Figure 5.18: Boxplot series of PLD, PTD, and PID over ODD. The data was grouped into ODD intervals of 500 metres whereby each interval's median and IQR are indicated by a black bar and a coloured box, respectively. Semi-transparent colors indicate that the respective interval contains less than 100 routes.

Figure 5.19 presents the correlation of PLD, PTD, and PID with ODD in given ODD intervals whereby the intervals were defined based on the results shown in Figure 5.18. Figure 5.19 confirms that the above-mentioned trends are significant. The results for San Francisco show that all three indices negatively correlate with ODD for ODDs between 6 and 9 kilometres and positively correlate with ODD when ODDs are above 9 kilometres. However, the effect sizes of the correlations are only weak to medium according to Cohen (1992). Regarding the trends in Shanghai, Figure 5.19 confirms weak positive correlation

for PTD and PID over almost all ODDs. In Vienna, correlations are also significant but generally very weak.

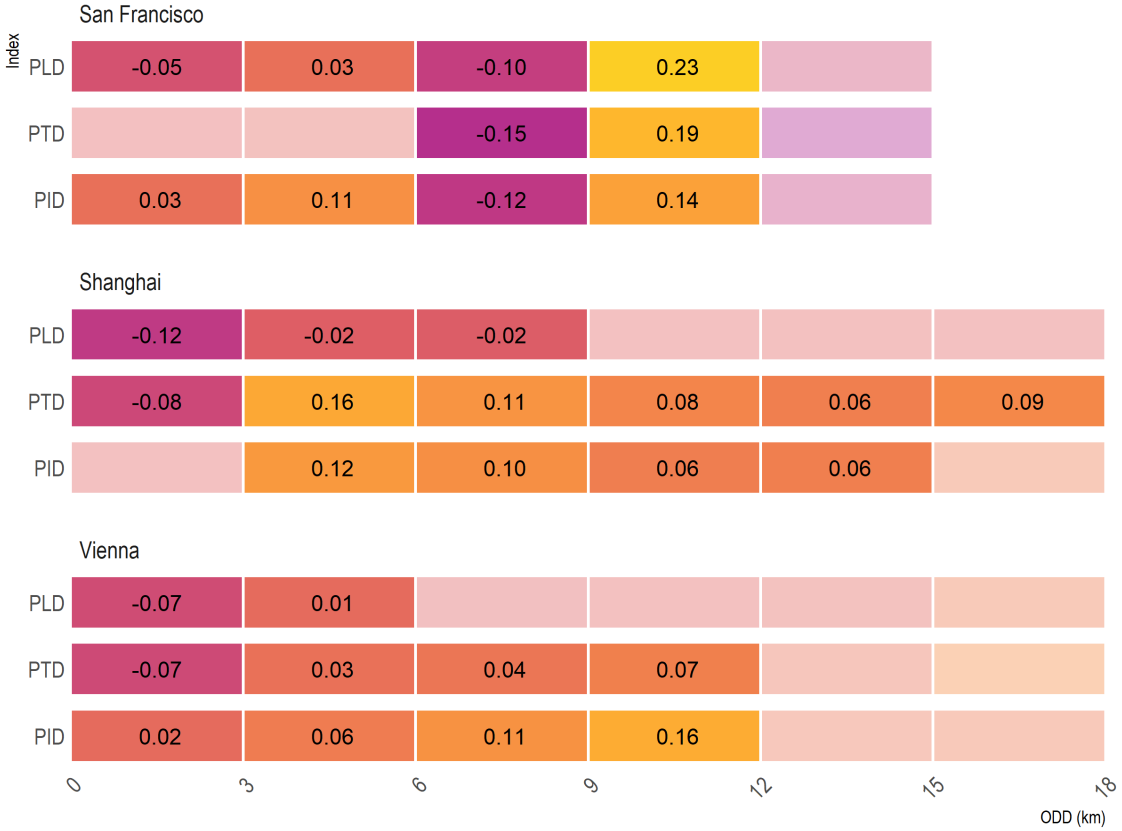


Figure 5.19: Correlation of PLD, PTD, and PID with ODD. The values represent the Spearman rank correlation coefficient ρ (Spearman 1904) which was used to calculate the correlation within each given ODD interval. According to Cohen (1992), $\rho = \pm 0.1$ corresponds to a weak effect, while $\rho = \pm 0.3$ implies a medium effect, and $\rho = \pm 0.5$ represents a strong effect. Dark values indicate negative and bright values indicate positive correlation. Semi-transparent tiles indicate that there is no significant correlation (p-value > 0.0001).

The results presented in this subsection can be summarised as follows:

General results

- The largest difference between actual and optimal routes in terms of length, duration, and number of intersections is presented by routes in San Francisco while taxi drivers' routes in Vienna differ the least from optimal routes.
- In all three cities, the difference between actual and optimal routes is smallest in terms of distance.

- The median actual route in San Francisco is 8.3 % longer than the shortest route while the difference is 9.9 % and 11.3 % in Shanghai and Vienna, respectively.
- 52 % of taxi drivers in San Francisco, 41 % of drivers in Shanghai, and 27 % of drivers in Vienna choose a route which is over 20 % slower than the fastest alternative.
- The proportion of actual routes including more than 10 % more intersections than the fewest intersections alternative lies between 50 % and 60 % in all three cities.

Results regarding a specific city

- In Shanghai and Vienna, the length difference between actual and shortest routes barely depends on the ODD.
- In Shanghai and Vienna, longer routes tend to be relatively slower and to include proportionally more intersections than the fastest and the fewest intersections alternatives.
- Taxi drivers' routes in San Francisco show a more complex relationship between the three route characteristics and ODD, namely that the taxi drivers' routes seem to perform better when they are longer.

For a discussion of these results, see Subsection 6.1.2.

5.3 Relationship between street network and routes

5.3.1 Centrality and locations of origins and destinations

This subsection visualises the spatial distributions of Edge Betweenness Centrality (EBC) as well as origin and destination locations. The results are then discussed in Subsection 6.2.1.

Figure 5.20 presents how EBC and actual routes are spatially distributed within the city's street networks. The maps in the left column show that edges with high BC are relatively evenly distributed in all three cities. The maps in the right column, however, show a strong spatial clustering of edges which are travelled by many actual routes. Although the maps do not appear to show similar patterns, and despite a pairwise comparison between the two maps of each city revealing that even in the city centres edges with high BC and edges travelled by many routes are often not congruent, there is significant positive correlation between an edge's BC and the proportion of actual trips traversing it (see Table 5.3). The values indicate weak correlation in San Francisco and Shanghai but strong correlation in Vienna.

Table 5.3: Correlation between EBC and share of actual trips per street network edge. The values were calculated using only edges travelled by at least one actual route. Correlation is represented using the Spearman rank correlation coefficient ρ (Spearman 1904). According to Cohen (1992), $\rho = \pm 0.1$ corresponds to a weak effect, while $\rho = \pm 0.3$ implies a medium effect, and $\rho = \pm 0.5$ represents a strong effect.

City	$\boldsymbol{\rho}$	**p**
San Francisco	0.11	$<$0.0001
Shanghai	0.16	$<$0.0001
Vienna	0.50	$<$0.0001

Figure 5.21 shows the spatial distribution of origin and destination locations. The maps reveal that the majority of trips is conducted between two locations in the city centre. This is the case in all three cities but most extreme in Shanghai where the two heatmaps are almost identical. Furthermore, it can be observed that in San Francisco and Vienna, there are more trips from the city centre to another area than vice versa as well as some trips using the highway connecting city centre and airport. The maps of San Francisco and Vienna further show clusters of origins and destinations which are located on the highway between the airport and the city centre. These clusters do not represent actual origin and destination locations but are caused cy the construction of actual routes from FCD (see Subsection 6.1.1 for further discussion). Unsurprisingly, there seems to be spatial correlation between the origin and destination locations in Figure 5.21 and the locations of the street segments passed by many routes shown in Figure 5.20.

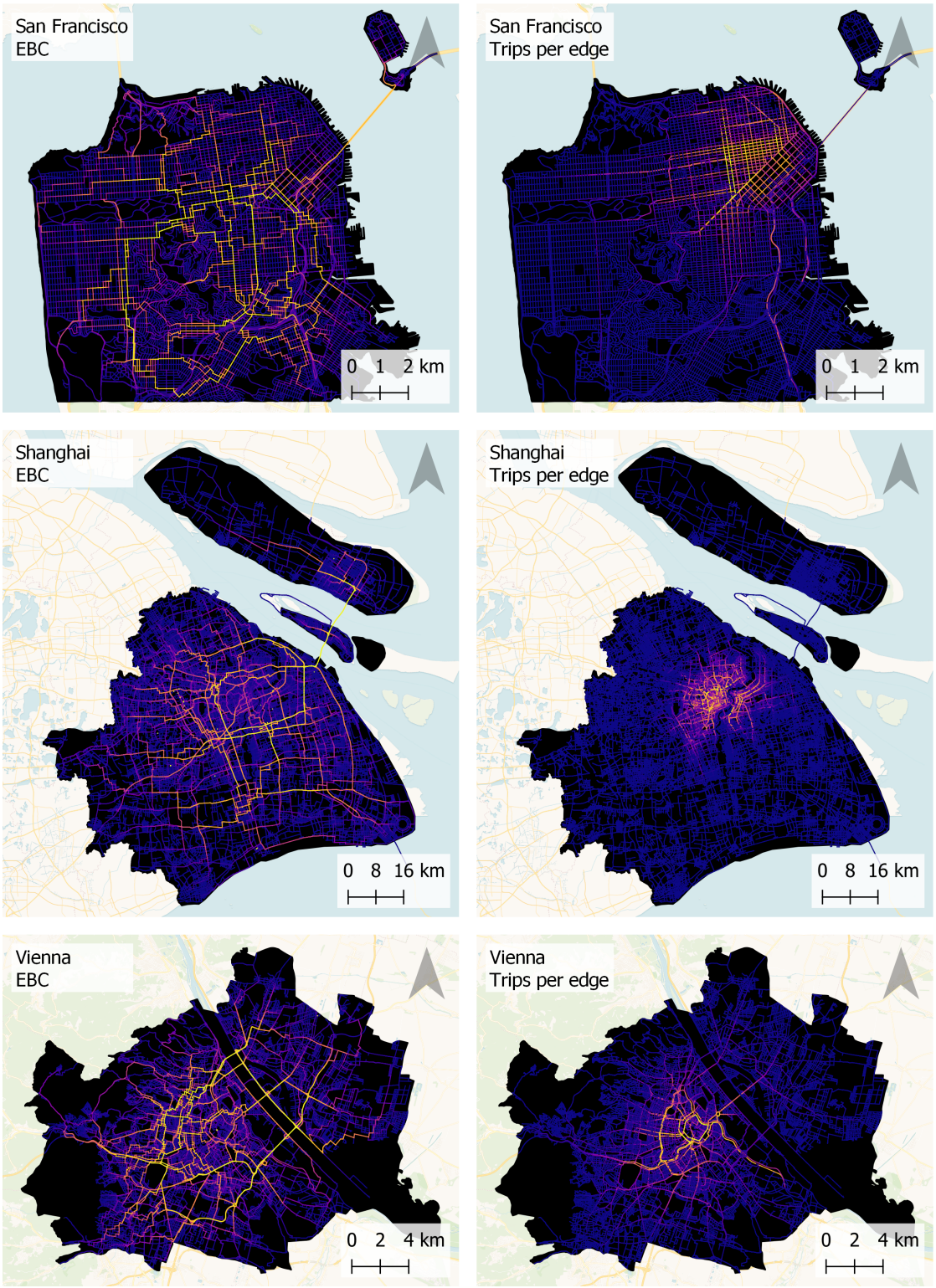


Figure 5.20: Spatial distribution of EBC and actual routes. For reasons of visualisation, the values were rescaled to an interval between [0,1]. Bright colors indicate high values.

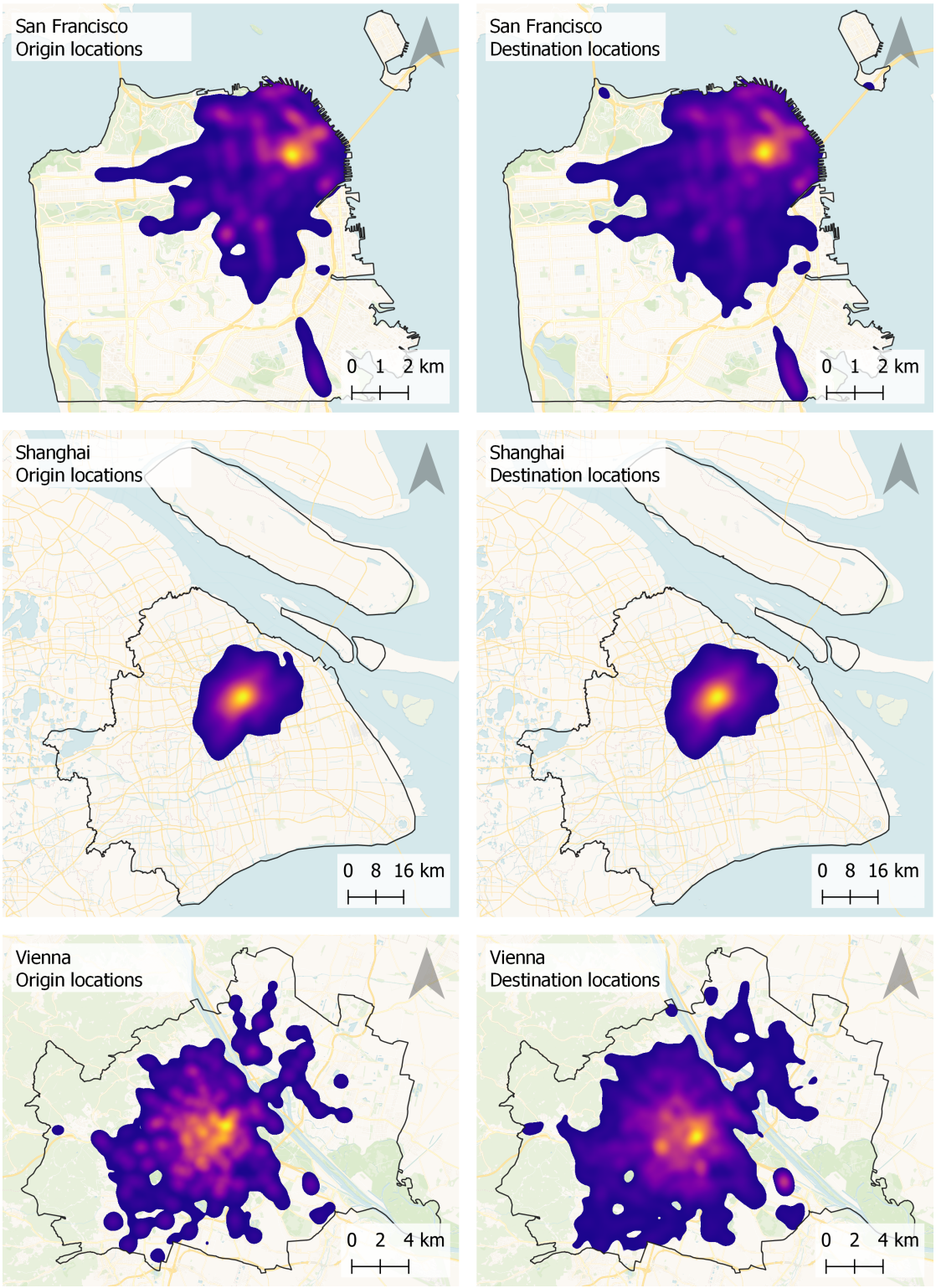


Figure 5.21: Spatial distribution of origin and destination locations. Bright colors indicate high values. Values in the 1st percentile are not shown.

The results presented in this subsection can be summarised as follows:

General results

- Street segments with high BC are spatially evenly distributed in all three cities.
- In all three cities, the majority of trips is conducted within the city centre.
- In all three cities, there is spatial correlation between the origin and destination locations and the locations of the most frequently driven streets.

Results regarding a specific city

- San Francisco and Shanghai present only weak positive correlation between an edge's BC and the proportion of actual trips traversing it.
- In Vienna, there is strong positive correlation between EBC and the proportion of trips per edge.

For a discussion of these results, see Subsection 6.2.1.

5.3.2 Road type composition

This subsection presents the results of the analysis of road types included in taxi drivers' routes. For a discussion of these results, see Subsection 6.2.2.

Figure 5.22 shows which fractions of each route type travel certain road types whereby the numerous OSM road types in the routes dataset were aggregated to a higher level distinguishing only between minor roads, major roads, and dual carriageways.[40]

[40] Minor roads include living and residential streets as well as all streets at the lowest network level and streets which do not fit into any other category. Major roads include primary, secondary, and tertiary roads which usually have directionally separated tracks. Dual carriageways include high performance roads where directions are usually physically separated. Additional information about the individual road types is presented in the Appendix.

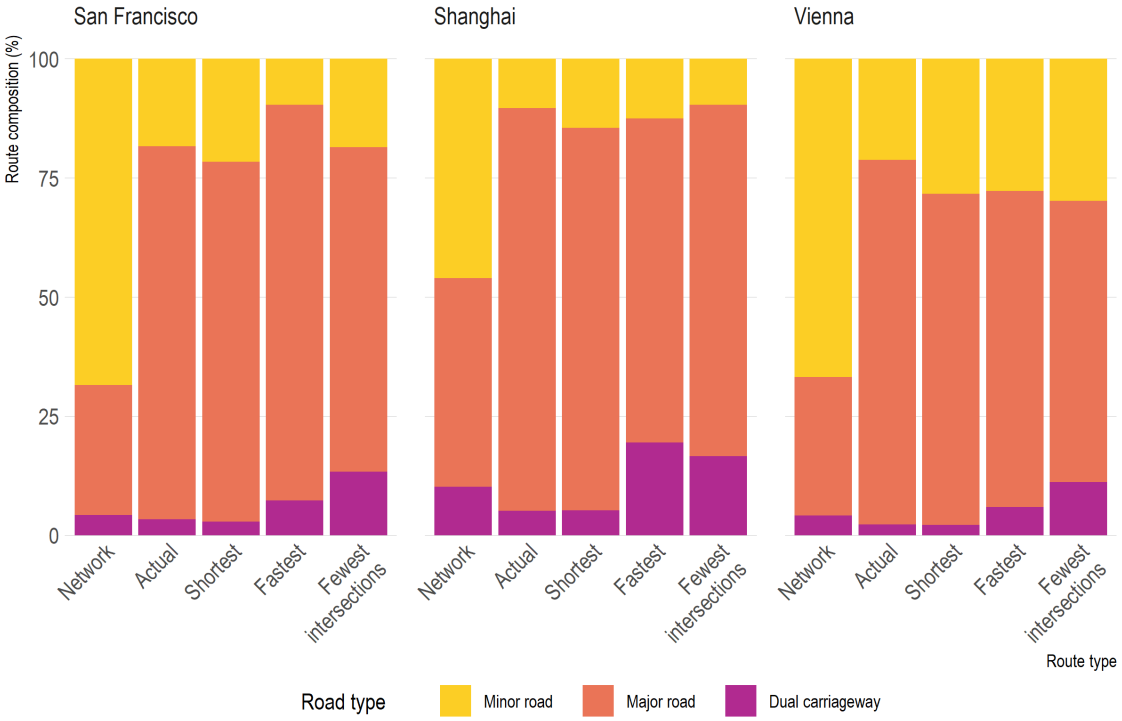


Figure 5.22: Composition of road types within each route type and in the street networks.

Figure 5.22 reveals that the largest proportion in all three street networks consists of minor roads and that dual carriageways only make up a small part of the total street length. However, they also reveal a difference between the almost identical road type compositions of the networks in San Francisco and Vienna, respectively, and the street network in Shanghai which presents a relatively small proportion of minor roads but larger proportions of major roads and dual carriageways. Looking at the results of the different route types, it stands out that in all three cities, all route types tend to include proportionally more major roads and less minor roads than the respective street network. Furthermore the fastest and fewest intersections routes in all three cities include larger proportions of dual carriageways than actual and shortest routes which even include proportionally less dual carriageways than the street network.

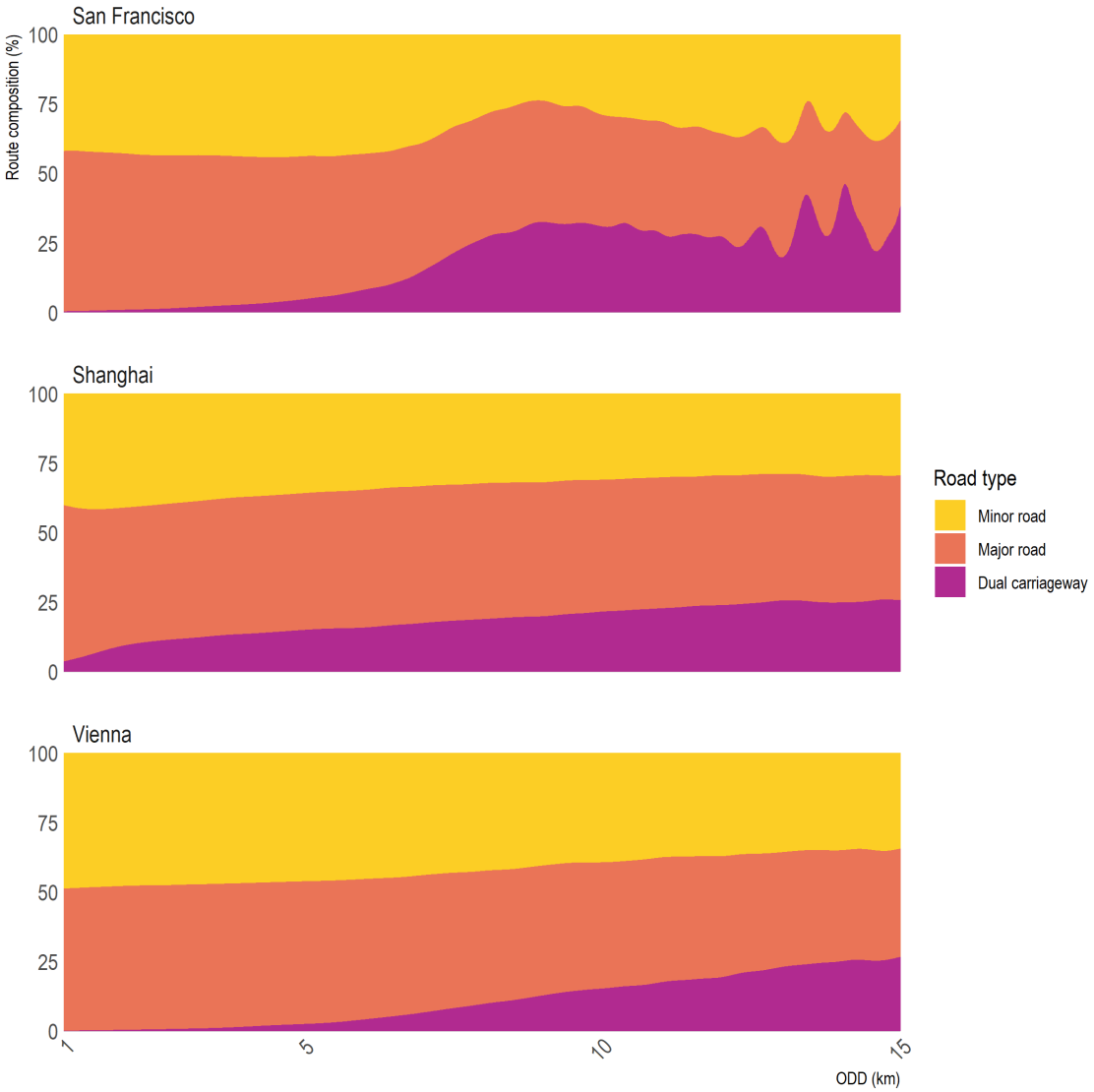


Figure 5.23: Composition of actual routes over ODD. Note that the San Francisco data are sparse in terms of routes with long ODDs and the results therefore need to be interpreted with caution.

Figure 5.23 presents insights about the relationship between actual routes' road type composition and Origin-Destination Distance (ODD). It reveals a general trend that the proportion of dual carriageways increases with increasing ODDs but also shows noticeable differences between the three cities. In San Francisco, routes with an ODD below 5 kilometres, which accounts for 87.7 % of the routes, consist of about 60 % major roads and about 35–40 % minor roads with only a very small proportion being dual carriageways. For routes with ODDs between 5 and 8 kilometres, the proportion of dual carriageways increases and the share of minor roads decreases with increasing ODDs while the pro-

portion of major roads remains roughly the same. Surprisingly, the trend seems to be exactly the other way around for routes with ODDs over 8 kilometres, although it must be said that these results are not very meaningful as only 2.7 % of the actual routes in San Francisco present such long ODDs. Shanghai presents a much more consistent trend of an increasing proportion of dual carriageways and a decreasing share of minor roads when ODDs become longer. The proportion of major roads stays about the same regardless of the ODD. It is worth mentioning that in Shanghai, even routes with short ODDs seem to travel on dual carriageways what makes it different from the other two cities. The Vienna plot shows the same overall trend as the Shanghai results with the difference that routes with ODDs below 5 kilometres almost exclusively use minor and major roads as it is the case in San Francisco.

Table 5.4 presents the effect size and significance of the correlation trends shown in Figure 5.23. It shows that in San Francisco, the positive correlation between the proportion of dual carriageways and ODD for routes with ODDs between 5 and 8 kilometres is the only trend showing more than only weak effect strength. The same goes for Shanghai where the only effect size worth mentioning is the medium positive correlation between the proportion of dual carriageways and the ODD. In Vienna, the correlations are generally stronger and also the correlation between the shares of minor and major roads, respectively, and the ODD is of medium strength.

Finally, Figure 5.24 visualises the spatial distribution of road types in comparison to the spatial distribution of actual routes. It confirms the finding from Figure 5.22 that the Shanghai street network contains proportionally less minor roads than the San Francisco and Vienna networks. Figure 5.24 also suggests that the city centres, where most of the actual routes are located, contain proportionally less minor roads than areas at the cities' peripheries.

Table 5.4: Correlation between proportions of road types in actual routes and ODD. The presented values represent the Spearman rank correlation coefficient ρ (Spearman 1904). According to Cohen (1992), $\rho = \pm 0.1$ corresponds to a weak effect, while $\rho = \pm 0.3$ implies a medium effect, and $\rho = \pm 0.5$ represents a strong effect. All presented values are highly significant (p-value < 0.0001). If no value is displayed (—), the result was not significant.

		ρ		
City	**ODD (km)**	**Minor road**	**Major road**	**Dual carriageway**
San Francisco	1 – 5	-0.02	—	0.10
San Francisco	5 – 8	-0.11	-0.19	0.29
San Francisco	8 – 15	0.09	0.11	-0.11
Shanghai	1 – 15	-0.03	-0.07	0.27
Vienna	1 – 15	-0.40	0.26	0.31

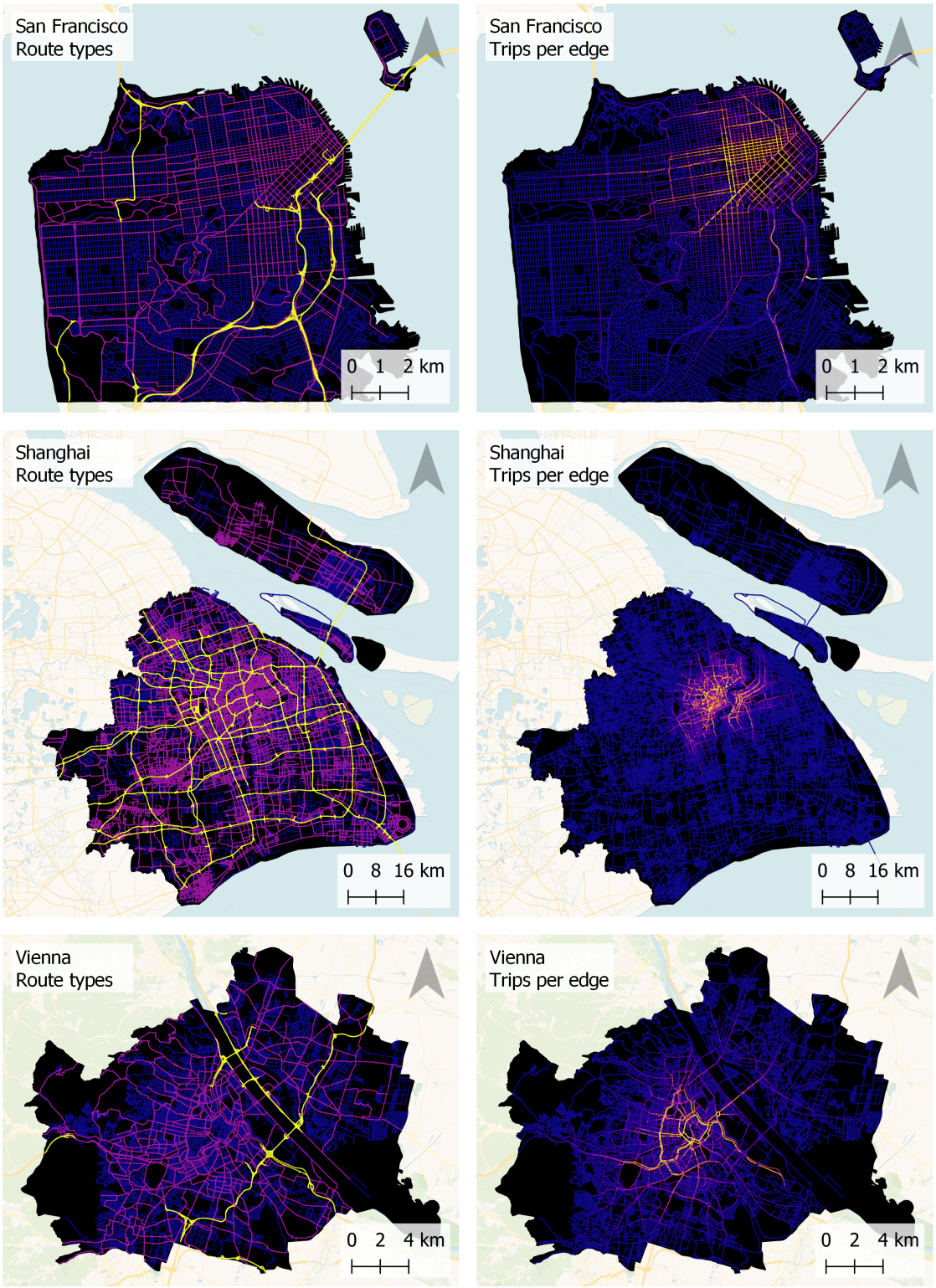


Figure 5.24: Spatial distribution of road types (left) displaying minor roads (blue), major roads (purple), and dual carriageways (yellow) and the proportion of actual routes per edge (right) whereby bright colors indicate high values.

The results presented in this subsection can be summarised as follows:

General results

- All three street networks consist mainly of minor roads.
- All route types in all cities tend to include proportionally more major roads and dual carriageways than the respective street network.
- Actual and shortest routes include proportionally less dual carriageways than the respective street networks, the fastest routes, and fewest intersections alternatives.
- All cities reveal a general trend that actual routes with longer ODDs include a larger share of dual carriageways but a smaller proportion of minor roads.
- City centres show lower proportions of minor roads and higher shares of major roads compared to the outer areas.

Results regarding a specific city

- The street network in Shanghai has proportionally more major streets and dual carriageways than the networks in San Francisco and Vienna.

For a discussion of these results, see Subsection 6.2.2.

5.3.3 Intersection density and complexity

This subsection presents the results from the analysis of intersections crossed by taxi drivers' routes. These results are then discussed in Subsection 6.2.3.

As already mentioned in Section 5.1, the differences in the routes' numbers of intersections shown in Figure 5.5 cannot be explained by the differences in intersection density presented in Table 3.1. However, Table 3.1 shows the number of intersections per area and not per street length which is somewhat counterintuitive as the drivers are bound to the street network when choosing their route. Figure 5.25 therefore presents the number of intersections per kilometre for each route type and compares the values to the street networks' intersection densities.

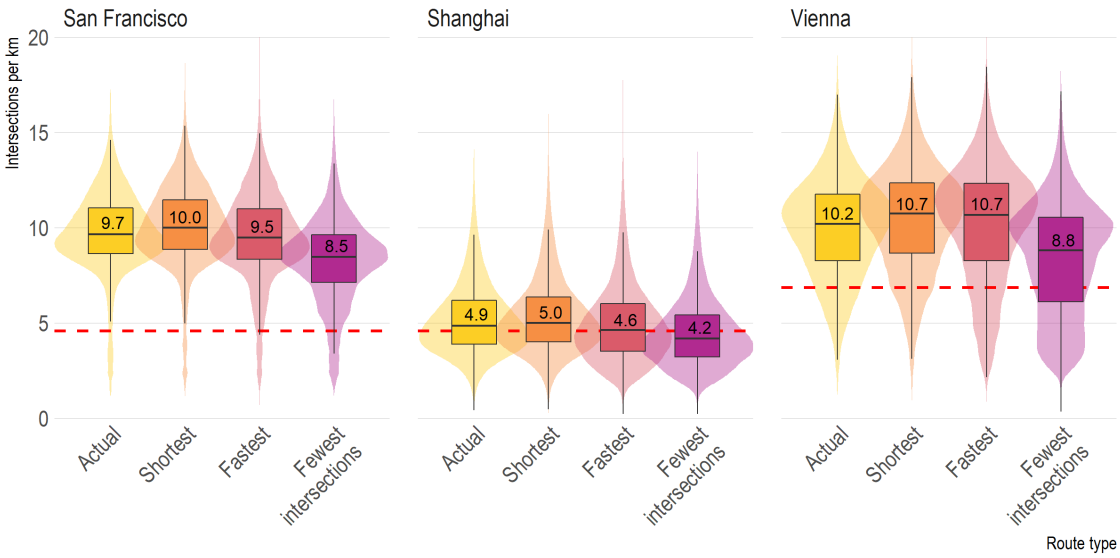


Figure 5.25: Number of intersections per kilometre of street length. The red dashed lines represent the street networks' intersection densities calculated from their total segment lengths and total numbers of intersections. The area behind each boxplot qualitatively represents the data distribution.

Overall, the results look similar to the numbers of intersections per route shown in Figure 5.5 because both plots present similar values for San Francisco and Vienna whereby routes in Shanghai cross less intersections. However, Figure 5.25 also shows the numbers of intersections per kilometre of street length in the whole street networks and reveals that Shanghai is the only city where the intersection density of the routes corresponds to the citywide intersection density. Routes in San Francisco include between 85 % and 117 % more intersections per kilometre then the average street kilometre in the network and routes in Vienna visit between 29 % and 57 % more intersections per kilometre than the network's average street. Comparing the route types within each city shows that shortest routes visit the most intersections per kilometre while fewest intersections routes present the lowest intersection density. Comparing the boxplots in Figure 5.25 to the bottom row of in boxplots in Figure 5.5 reveals that actual and fewest intersections routes perform better in terms of intersection density than in terms of total number of intersections when compared to shortest and fastest routes.
Figure 5.26 presents a comparison of the different route types and the street networks in terms of the complexity of the intersections they visit.

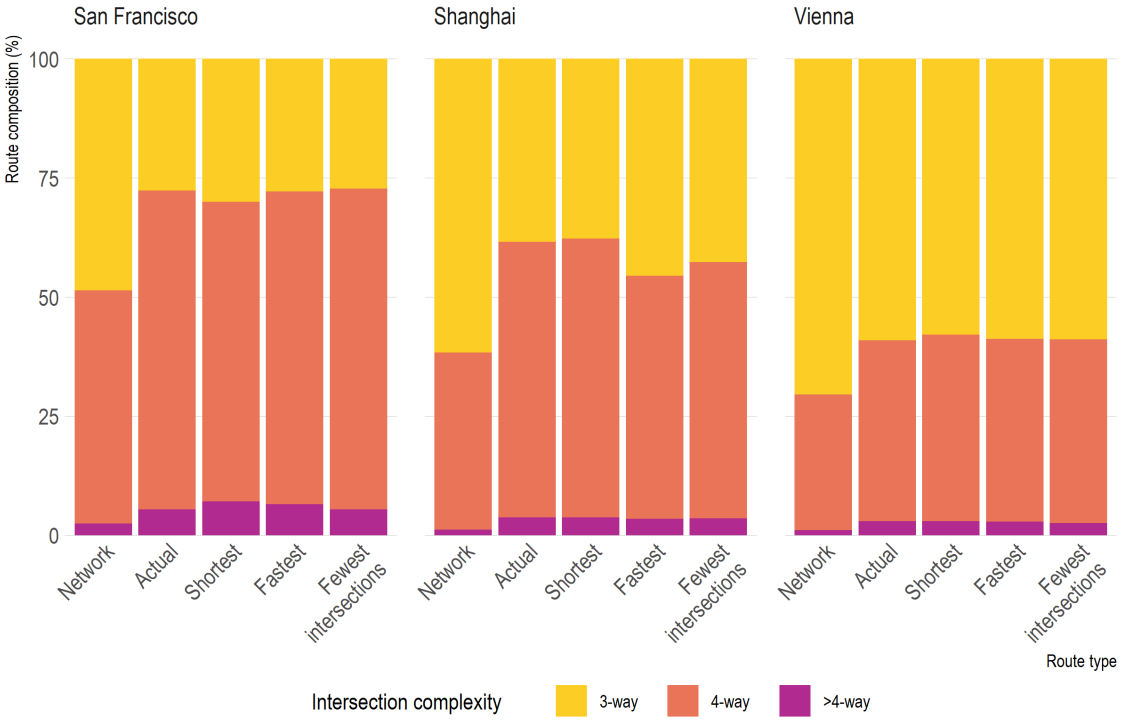


Figure 5.26: Composition of intersections' complexity within each route type and in the street networks.

It is apparent that the largest part of all intersections, in the street networks as well as in the routes, are 3-way and 4-way intersections. In all three cities, intersections with more than 4 streets make up less than 2.5 % of the intersections in the street network but are proportionally overrepresented in all route types. Comparing the street networks' and the routes' proportions of 3-way and 4-way intersections shows the same pattern in all three cities, namely that all route types contain proportionally more 4-way intersections than the respective road network. The relative difference is largest in Shanghai where on average, a routes' proportion of 4-way intersections is 49 % higher than the street network's. In San Francisco and Vienna, the routes' proportion is 34 % and 35 % higher, respectively. Having a look at the results within each city reveals no major differences between the route types. Figure 5.27 presents the spatial distributions of intersections in all three cities. The left column of maps shows that the intersection density in San Francisco is similar in all parts of the city while in Vienna and Shanghai, intersections are clustered in the city centre. A comparison of the spatial distributions of all intersections with the distribution of complex intersections[41] shows that they are similarly distributed in all three cities.

[41] Based on the results presented in Figure 5.26, a complex intersection is defined as an intersection with 4 or more streets.

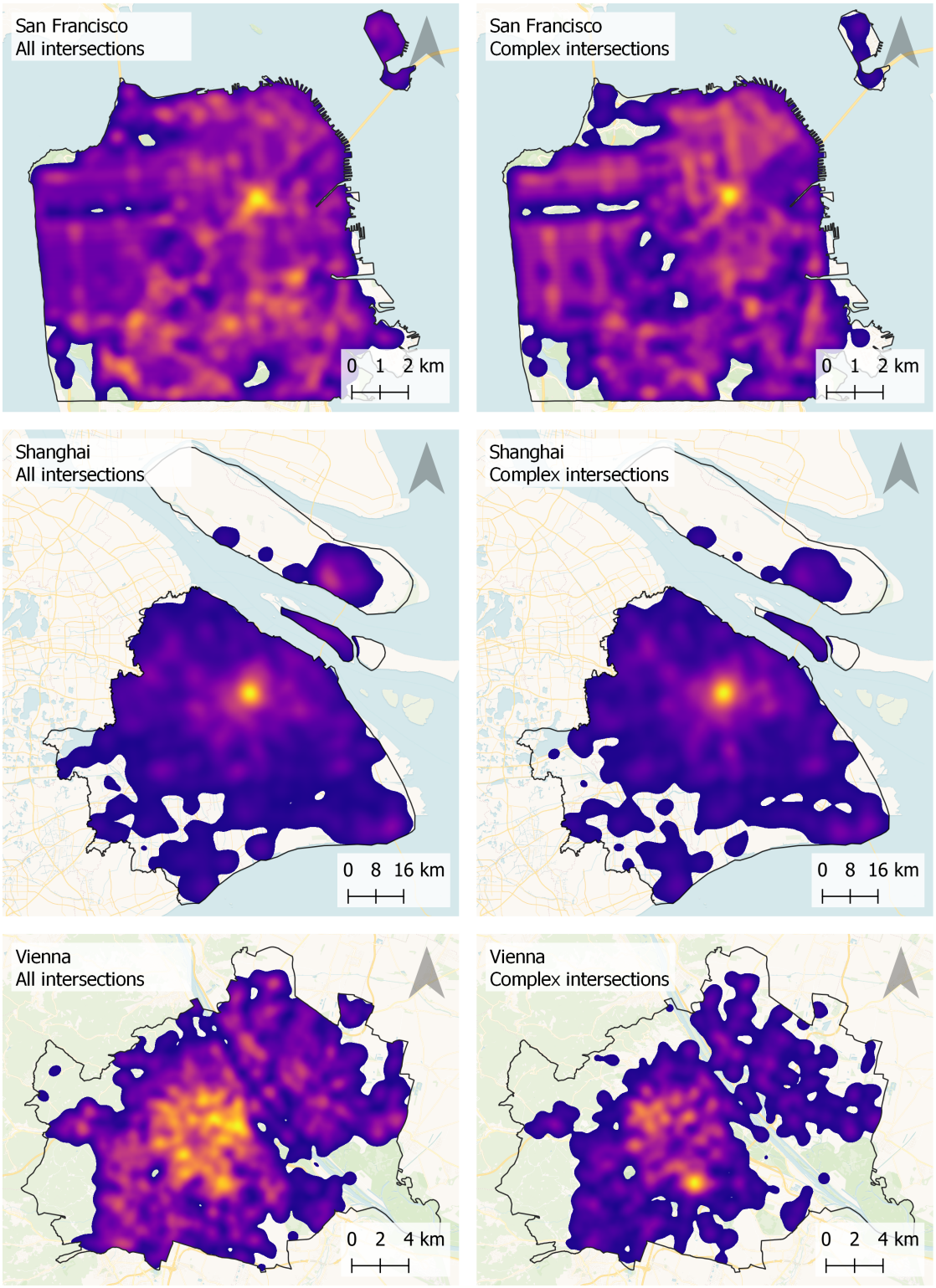


Figure 5.27: Spatial distribution of all intersections (left) and of intersections with more than 3 streets (right). Bright colors indicate high densities of intersections. Values in the 1st percentile are not shown.

The results presented in this subsection can be summarised as follows:

General results

- The number of intersections a route visits per kilometre differs between the cities.
- Actual routes compare better with optimal routes in terms of intersections per kilometre than in terms of total number of intersections. They present similar intersection densities as shortest and fastest routes.
- Most route types show a higher proportion of complex intersections than the respective street network.
- There are only minor differences between the route types within each city in terms of intersection complexity.

Results regarding a specific city

- There is a discrepancy between the intersection densities of routes and street networks in San Francisco and Vienna where routes show higher intersection densities than the networks.
- In San Francisco, the spatial distribution of intersections is similar in all city districts.
- In Shanghai and Vienna, the intersection density in the city centre is higher than in the peripheral areas.

For a discussion of these results, see Subsection 6.2.3.

5.3.4 Number of turns and turn characteristics

This subsection presents the results from the analysis of turns in taxi drivers' routes. The results are then further discussed in Subsection 6.2.4.

Figure 5.28 presents the number of turns for each route type. It can be observed that routes in Shanghai show the lowest numbers of turns while they are slightly higher in San Francisco and highest in Vienna. The differences between the three cities are small as all median values lie between 3 and 6 turns per route. However, the ranges of values and the IQRs differ with Vienna showing the largest variation.

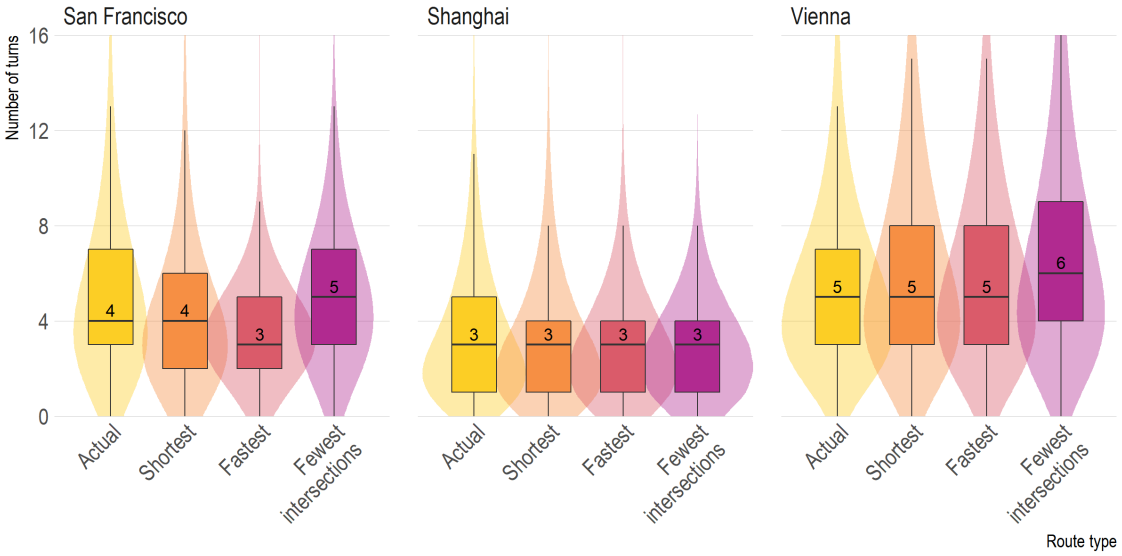


Figure 5.28: Total number of turns per route. The area behind each boxplot qualitatively represents the data distribution. Note that the areas represent discrete data but were smoothed for better visualisation.

Figure 5.28 reveals another noticeable result as it shows that the fewest intersections routes in San Francisco and Vienna tend to include more turns than actual, shortest, and fastest routes. Since intersections and turns are strongly connected, this finding is somewhat counterintuitive and shall therefore be addressed here in order to avoid misunderstandings about the remaining findings. For this purpose, it is important to understand the difference between the definitions of intersections and turns used in this thesis: an intersection is a location where 3 or more streets meet (see Section 4.6) and the numbers of intersections a route visits (see Figure 5.5) is simply the number of such locations passed. However, as passing an intersection does not require the driver to make a turn, a route can include many intersections without making any turns. A turn on the other hand is defined as as a change of direction where the angle of directional change is larger than 45 degrees (see Section 4.6 and Figure 5.2) whereby turns can only be made at intersections. This means that while a route cannot contain less intersections than turns, it can contain less turns than intersections which is the case for all route types (see figures 5.5 and 5.28). The fact that fewest intersections routes in San Francisco and Vienna contain more turns than the other route types can therefore be explained by assuming that fewest intersections routes make proportionally more turns, meaning that although they visit less intersections, they use proportionally more of them to make a turn. This hypothesis is supported by Figure 5.29 which shows that the proportion of intersections at which a turn is made is highest for fewest intersections routes in San Francisco and Vienna while it is similar to the other route types' results in Shanghai.

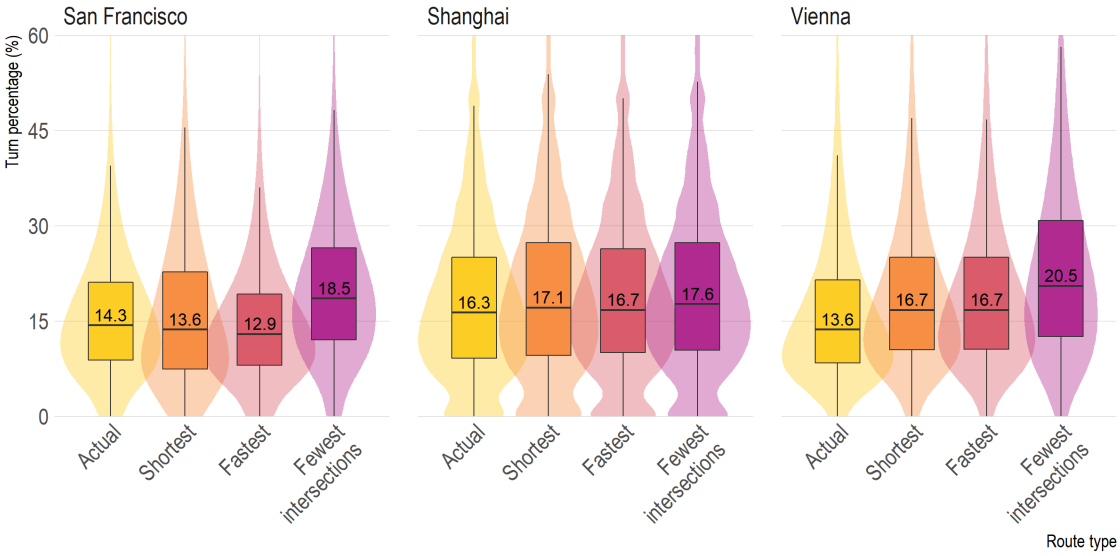


Figure 5.29: Percentage of intersections at which a route makes a turn. The area behind each boxplot qualitatively represents the data distribution.

Figure 5.29 further shows that fewest intersections routes have the highest turn percentage in all three cities and that actual routes have the lowest turn percentage in Shanghai and Vienna. Furthermore, the plots show that the values' distributions are similar in all three cities and for all route types with the exception that all route types in Shanghai seem to include relatively many routes with 0 turns.

Figure 5.30 presents the proportion of turns with different characteristics. It reveals similar proportions for all route types in all three cities and suggests that, regardless of the route type, right and left turns occur equally often and that about half of all turns are sharp turns.

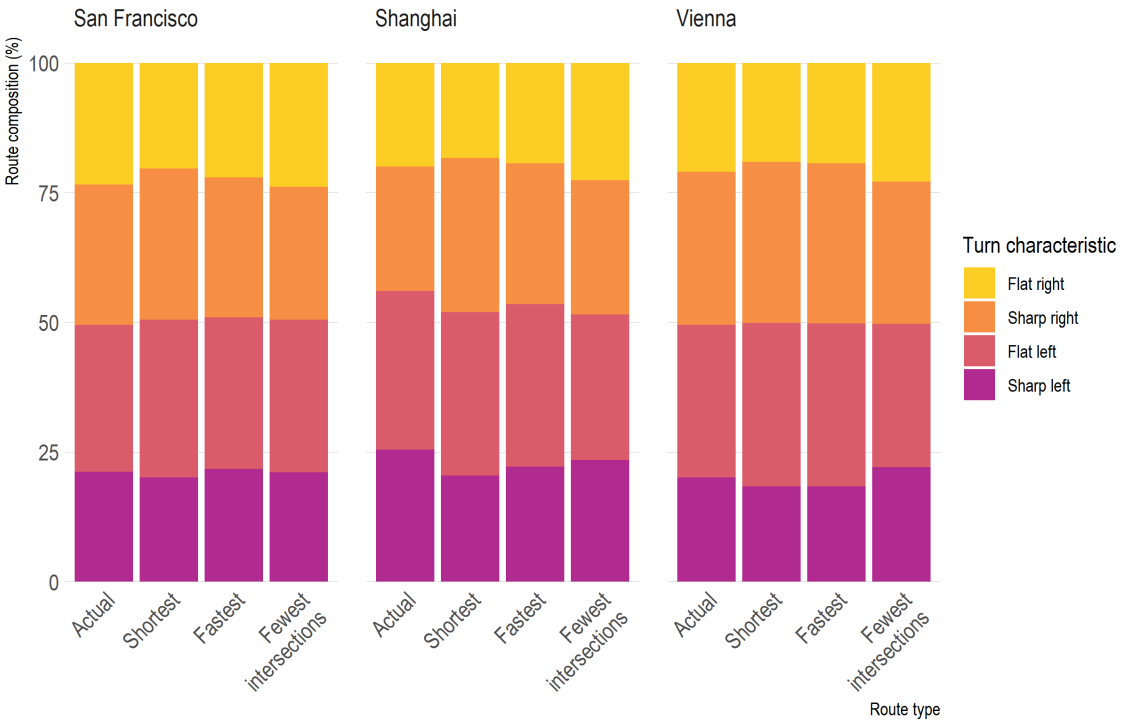


Figure 5.30: Proportions of flat right, sharp right, flat left, and sharp left turns for each route type. For a definition of flat and sharp turns, see Section 4.6.

The results presented in this subsection can be summarised as follows:

General results

- The median numbers of turns per route slightly differ between the three cities with routes in Shanghai including the fewest and routes in Vienna including the most turns.
- In all three cities, actual routes can compete with shortest and fastest routes in terms of number of turns.
- In all three cities, fewest intersections routes present the highest turn percentage.
- Right and left turns as well as flat and sharp turns occur about equally often in all cities and all route types.

Results regarding a specific city

- In San Francisco and Vienna, fewest intersection routes include the most turns.
- In San Francisco and Vienna, actual routes show the lowest turn percentage.
- Shanghai seems to include many routes with 0 turns.

For a discussion of these results, see Subsection 6.2.4.

Chapter 6

Discussion and synthesis

6.1 Similarity between actual and optimal routes

6.1.1 Percentage of shared length and Fréchet distance

This subsection discusses the results from the analysis of route similarity. The results were presented in subsections 5.2.1 and 5.2.2.

It has been revealed that the extent to which taxi drivers' routes follow shortest, fastest, and fewest intersections routes differ between the cities as well as between the individual routes. This finding is based on the differences and variation in the routes' PSL presented in Figure 5.6. However, the results of the analysis of the Fréchet Distance (FD), which are presented in Figure 5.14, suggest the opposite, namely that the extent to which actual and optimal routes are alike is similar between the three cities as well as between individual routes. These two statements seem contradictory, which is due to the fact that PSL and FD are both used as measures of route similarity although they do not measure the same thing. The PSL defines similarity by the proportion of a route that exactly overlaps with an optimal route, which means that two routes are only considered similar if they match exactly. The FD defines similarity by the spatial distance between two routes and therefore catches similarity even when the routes are close but not congruent. However, a small FD does not automatically mean that the two routes run on the same streets. Thus, the PSL measures the extent to which two routes are identical while the FD quantifies the extent to which two routes are geometrically similar. It can therefore be said that although the extent to which the taxi drivers' routes and optimal routes are similar is comparable in all three cities, there are differences between the cities with regard to the extent to which the drivers follow the optimal routes. Regarding the individual routes in each city, it can be said that while all drivers tend to choose a route that is similar to an optimal route, not all drivers follow it equally consistently.

The results presented in Subsection 5.2.1 have shown that in general, taxi drivers in Shanghai are most likely to follow an optimal route while drivers in San Francisco are least likely to follow optimal routes. However, as the extent to which their routes follow optimal routes shows no major differences, these different behaviours have only little influence on the length, duration, and number of intersections of the daily driven routes. Furthermore, the comparison of figures 5.6 and 5.14 reveals that although the routes in San Francisco present lower PSL values, their FDs are comparable to the FDs of routes in the other two cities. This pattern can be explained by assuming that drivers in San Francisco use parallel streets more often than their colleagues in Shanghai or Vienna. This would also make sense insofar as the road network in San Francisco is fundamentally different from the road networks in Shanghai and Vienna (see Figures 3.1, 3.2, and 3.3) as it is laid out in a regular, rectangular grid with many parallel streets.
Figures 5.7 and 5.8 have suggested that taxi drivers in San Francisco are more likely to follow optimal routes when their trip is very long than when it is only a few kilometres. However, this seems odd as this trend is very different from the results from Shanghai and Vienna. Figure 6.1 visualises the spatial distribution of routes with Origin-Destination Distances (ODDs) longer than 6 kilometres as these routes are the ones causing the mentioned inconsistency in the PSL trend.

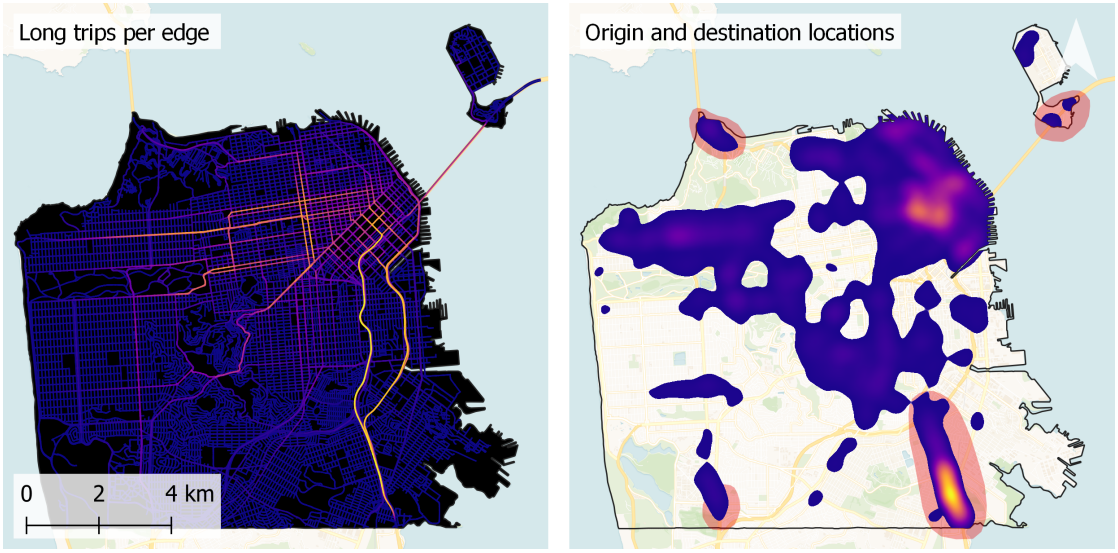


Figure 6.1: Spatial distribution of long actual routes and their origin and destination locations. Here, long routes are defined as routes whose ODD is longer than 6 kilometres. Brighter colors indicate higher values. In the plot on the right, the red areas indicate clusters caused by the truncation of routes. Furthermore, values in the 1st percentile are not shown.

The map on the left reveals that a large share of these long routes run on the highway connecting the city centre and the airport. Since the highway includes no intersections and is obviously the shortest and fastest way to cross the city in a north-south direction, these routes present a high PSL which in turn explains the increase in PSL presented in Figure 5.7. The second map presented in Figure 6.1 shows the spatial distribution of origin and destination locations of routes with ODDs above 6 kilometres. The marked areas indicate clusters of trip origins or destinations which are likely due to the truncation of trips. It can be clearly seen that the most extreme cluster is located on the highway leading to the airport. As already mentioned in Subsection 5.3.1, these origin and destination locations do not reflect the true start and endpoints of the trips but are the consequence of the chosen FCD processing approach (see Section 4.4). The trips between airport and city were truncated during the construction of the routes from the FCD because data points located outside of the city's boundaries were removed before the data were segmented into individual trips. This had the effect that the data points representing the real origin and destination locations at the airport as well as the paths between the airport and the city boundary were not available for the construction of the actual routes. In retrospect, the FCD should have been processed differently in order to avoid this truncation of routes. The preferred alternative to the taken approach would be to construct the actual routes based on all data points in the dataset and then remove entire routes if they intersect the city boundaries. Of course, any outliers should be removed before the routes are created but the intersection with the area of interest should only be done with the complete routes and not with the individual data points.[42] In order to assess the influence of truncated trips on the results presented in Figure 5.7, the San Francisco boxplot series are plotted again after all routes where the origin or destination location lies within one of the areas marked in Figure 6.1 were removed. The result is presented in Figure 6.2 and although the middle plot still shows an intermediate increase in PSL, the overall trend is clearly a decrease in PSL with increasing ODD. The truncated routes thus appear to clearly influence the analysis of the PSL and will therefore be disregarded in the further discussion. To a lesser extent, this also applies to the long routes in Vienna since these routes also often run between the city and the airport and were truncated during route construction (see Subsection 5.3.1). To return to the above-mentioned increase in PSL on long routes, it can be said that this trend does not reflect the actual route choice behaviour of taxi drivers but rather the fact that a large part of the long routes in San Francisco runs between the city centre and the airport and that there are no sensible alternatives for this route.

[42] Including the routes to the airport would also be a possibility but this would mean an adjustment of the area of interest.

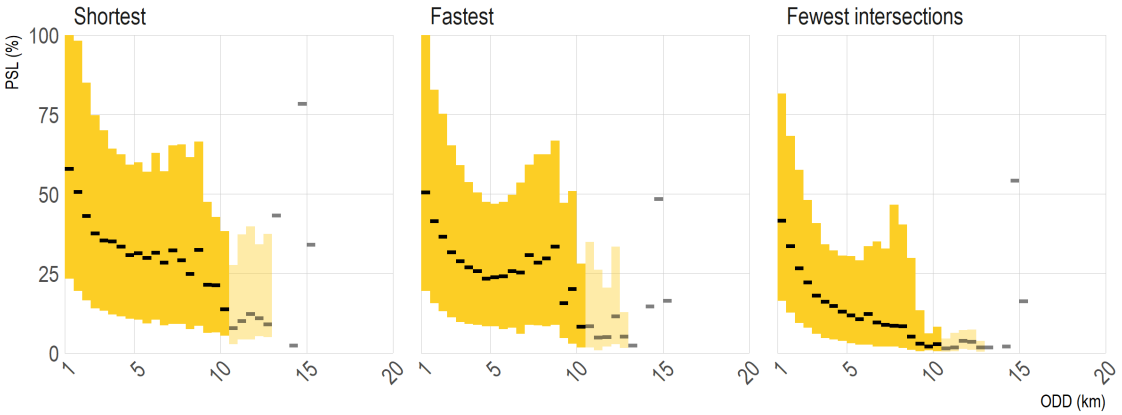


Figure 6.2: Boxplot series of PSL over ODD in San Francisco. Routes which are likely to be truncated were removed. The data was grouped into ODD intervals of 500 metres whereby each interval's median and IQR are indicated by a black bar and a coloured box, respectively. Semi-transparent colors indicate that the respective interval contains less than 100 routes.

Another finding visualised by figures 5.7 and 5.15 is that the drivers in Shanghai and Vienna tend to prefer shortest routes over fastest and fewest intersections routes, which is particularly the case when trips are long. This suggests that taxi drivers in Shanghai and Vienna attach more importance to the length of their route when their trip is long. Drivers from San Francisco, however, do not show this behaviour as they are most likely to optimise for distance on short and long routes.

The results presented in figures 5.9 and 5.13 have revealed that there is no difference between daytime and nighttime routes in San Francisco and Shanghai but that taxi drivers in Vienna tend to choose routes which are less similar to optimal routes during the night when the route's ODD is longer than 6 kilometres. As this very specific difference can be observed to the same extent for fastest, shortest, and fewest intersections routes, it is probably caused by the fact that in Vienna certain routes are not driven at night. Figure 6.3 therefore shows the spatial distribution of actual routes with ODDs longer than 6 kilometres during the day and during the night.

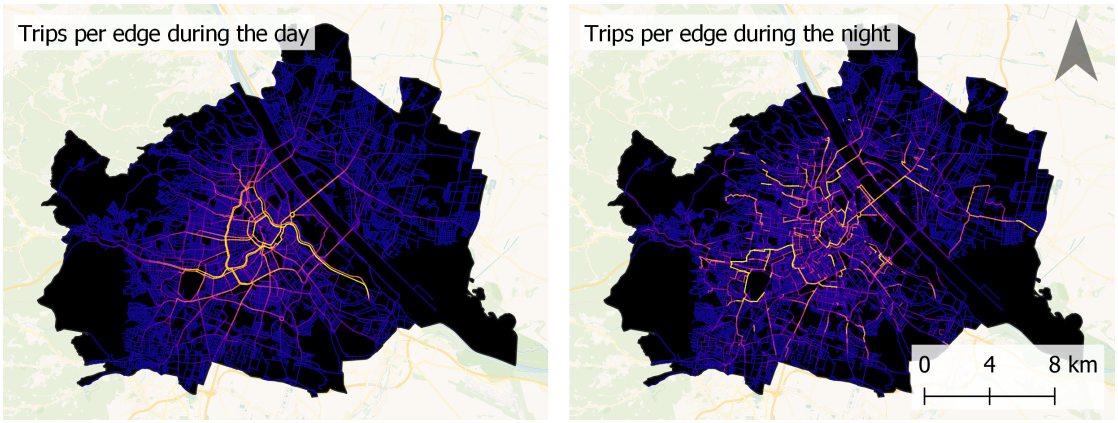


Figure 6.3: Spatial distribution of long actual routes during day and night in Vienna. Here, long routes are defined as routes whose ODD is longer than 6 kilometres. Brighter colors indicate higher values.

The two maps in Figure 6.3 reveal that the spatial distribution of long routes in Vienna during the day and at night is very different. During the day, most of the routes run along the main traffic axes in Vienna's road network, namely the so-called "Hauptstrasse B" and on the Autobahn (City of Vienna 2012). At night, however, the routes are more spread out into the residential areas where they run on low-ranking roads. The finely branched street network in the residential areas provides more parallel roads and alternative routes than the major roads around the city centre which explains why the night routes have a lower PSL. The absence of significant differences between day and night in San Francisco and Shanghai indicates that neither the spatial distribution of the routes nor the route choice behaviour of the taxi drivers differs significantly between day and night.
Figures 5.9 and 5.11 have shown that taxi drivers in San Francisco are less likely to follow the shortest, fastest, and fewest intersections routes on weekends. Figure 6.4 presents a visual comparison of the spatial distribution of routes during the week and on the weekend and reveals that during the week, a higher proportion of the routes is using the highway. Since the highway is more likely to be congruent with the shortest, fastest, and fewest intersections routes, this finding is able to explain the difference in PSL between routes during the week and on the weekends. The reason for the different utilisation of the highway by taxis may be the lack of commuter traffic to and from the city centre during the weekend.

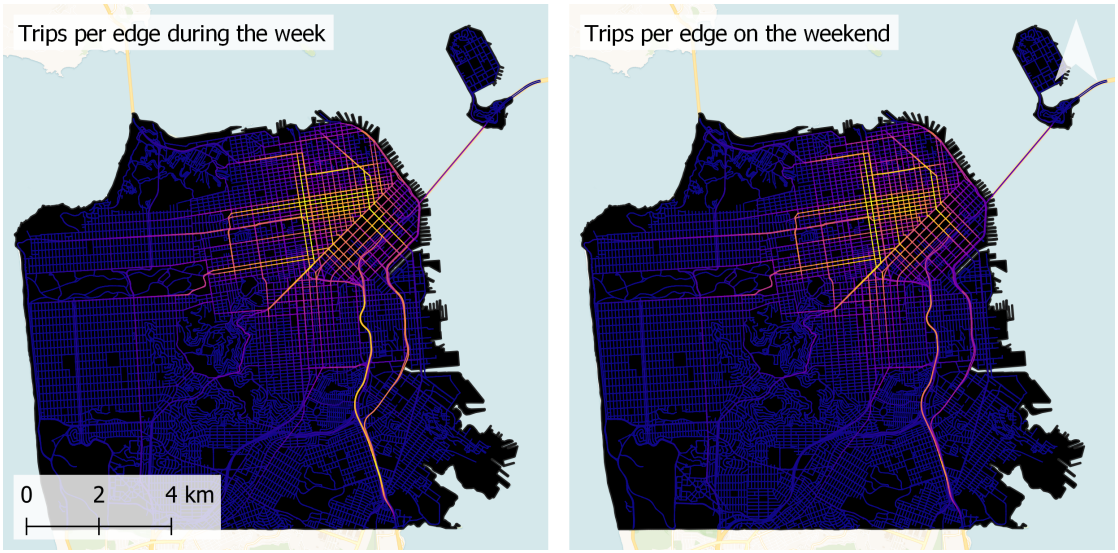


Figure 6.4: Spatial distribution of actual routes during the week and on weekends in San Francisco. Brighter colors indicate higher values. Note that the two color scales are independent.

The analysis of route similarity has revealed the following findings:

General findings

- Although all drivers tend to choose a route which is similar to an optimal route, the extent to which the individual routes follow the shortest, fastest, or fewest intersections alternatives vary.

Findings regarding a specific city

- In San Francisco, only 18.1 % of the drivers choose the shortest, 13.5 % the fastest, and 9.6 % the fewest intersections route.
- In Shanghai, 38.8 % of the drivers follow the shortest, 33.9 % the fastest and 34.0 % the fewest intersections route.
- In Vienna, 25.6 % of the drivers choose the shortest, 24.2 % the fastest, and 16.9 % the fewest intersections route.
- In San Francisco, routes with an ODD above 6 kilometres tend to be more congruent with optimal routes than when the ODD is only a few kilometres. This trend is most likely caused by the majority of these long routes connecting the city centre and the airport which is a route without reasonable alternatives.
- Taxi drivers in Shanghai and Vienna attach more importance to the length of their route when their trip is long while drivers from San Francisco are most likely to optimise for distance on short and long routes.

- In Vienna, long routes chosen during the day are more congruent with optimal routes than long routes taxi drivers choose at night.
- In San Francisco, routes conducted on weekends are less similar to optimal routes.

6.1.2 Percentage difference of length, duration, and number of intersections

This subsection discusses the results from the analysis of Percentage of Length Difference (PLD), Percentage of Time Difference (PTD), and Percentage of Intersections Difference (PID). The results were presented in Subsection 5.2.3.

The results have shown that the routes chosen by taxi drivers in the three cities do not perform equally when compared with the shortest, fastest, or fewest intersections route. For example, the general difference between actual and optimal routes is smallest in Vienna and largest in San Francisco while the routes from Shanghai are somewhere in between. Furthermore, the routes in Vienna also show fewer differences, indicating that the route choice behaviour of drivers in Vienna is more homogeneous than that of their colleagues in San Francisco and Shanghai. What is the same in all three cities is that the difference between actual and optimal routes is smallest in terms of length. In terms of duration, the routes in San Francisco and Shanghai show large deviations from the fastest route with over half of the drivers in San Francisco and over 40 % of the drivers in Shanghai choosing a route that is more than 20 % slower than the fastest alternative. In Vienna this share is significantly lower at only 27 %. In terms of the number of intersections, the routes in San Francisco again show the most deviation from fewest intersections routes while the routes in Vienna again perform best. However, the differences here are relatively small with the proportion of routes including more than 10 % more intersections than the fewest intersections alternative lying between 50 % and 60 % in all three cities.
In addition to the results mentioned above, figures 5.18 and 5.19 have shown that the relative difference between actual and optimal routes in the three cities depends to varying extents on Origin-Destination Distance (ODD). Probably the most surprising result is presented by San Francisco, namely that routes with an ODD between 6 and 9 kilometres tend to perform better than shorter routes in terms of length, duration, and number of intersections. This distinguishes the routes in San Francisco from those in Shanghai and Vienna which show similar trends for all three characteristics as in both cities, the routes chosen by taxi drivers do not seem to deviate more or less from the shortest route if they have long ODDs but the length difference between actual and shortest route is constant for long and short trips. In addition, the relative difference in terms of duration

and number of intersections steadily increases as ODD increases in both cities. The PID results from Vienna present a similar pattern as the results from San Francisco as the PID drops at ODDs of around 14 kilometres. However, less than 1 % of the routes from Vienna have an ODD longer than 13 kilometres and therefore, this finding is not further discussed. The pattern revealed by the results from San Francisco, however, seems more solid as it is visible in all three plots and supported by 6.9 % of the actual routes. A possible explanation might be that because the city is relatively small, many routes with such long ODDs are travelling via the highway. Assuming that taxi drivers also tend to use the highway for long routes, these long routes would be more congruent with the optimal alternatives. This theory is supported by the fact the proportion of dual carriageways in actual routes is highest when ODDs are between 7 and 10 kilometres (see Figure 5.23) but it does not explain why the PLD also shows a decrease in the given ODD range.

The fact that actual routes in all three cities differ least in length from the optimal alternative suggests that taxi drivers are most likely to optimise their route choice based on distance. This seems to be the case especially for long routes as the PLD is the index showing the least correlation with the ODD. In San Francisco, the PTD also changes only slightly as the trips get longer, but the relative difference between the duration of the actual route and the duration of the fastest route is generally about twice as high as the length difference between actual and shortest route which suggests that taxi drivers prefer a short route to a fast route. In Shanghai and Vienna, this seems to be the case as well as the PTD increases with increasing trip length. Also, the number of intersections does not seem to be a priority in any of the cities since the relative difference between actual and fewest intersections routes is larger than the relative length difference between actual and shortest route in all cities and for all ODDs.

Although the results have revealed that there are differences between the cities when it comes to comparing taxi drivers' routes with optimal routes, these differences are rather small. Assuming that one would book a trip whose optimal route's length, duration, and number of intersections correspond exactly to the median of the whole dataset, the optimal route would be 3.41 kilometres long, take 5.7 minutes, and cross 21 intersections. A hypothetical median Viennese taxi driver would then choose a route which is 3.69 kilometres long, takes 6.3 minutes and includes 23.2 intersections while the route chosen by a driver from Shanghai would be 3.75 kilometres in length, 6.6 minutes in duration, and crossing 23.5 intersections. The route chosen by the hypothetical taxi driver from San Francisco would be 3.79 kilometres long, 6.9 minutes long, and including 23.9 intersections. This thought experiment shows, that the difference between the cities are small as the routes chosen by three different drivers from the three cities would only differ by 102

metres in length, 39 seconds in duration, and less than 1 intersection. The fact that the number of intersections differs the least might also be a reason why it is not prioritised as a route choice criterion by the taxi drivers.

The analysis of how taxi drivers' routes compare to optimal routes in terms of length, duration, and number of intersections has revealed the following findings:

General findings

- Although the relative differences between the cities seem large, the actual differences in terms of length, duration, and number of intersections are small when the numbers are applied to taxi drivers' actual routes.
- Taxi drivers in all three cities tend to prioritise distance when choosing a route.

Findings regarding a specific city

- The routes chosen by taxi drivers in Vienna differ the least from optimal routes in terms of length, duration, and number of intersections.
- In San Francisco, the extent to which a route differs from the optimal routes in terms of length, duration, and number of intersections is depending on the length of the trip.
- In Shanghai and Vienna, the relative length difference between actual and shortest route does not change when a route is longer or shorter. However, longer routes tend to present a higher relative difference to the optimal routes in terms of duration and number of intersections.

6.2 Relationship between street network and routes

6.2.1 Centrality and locations of origins and destinations

This subsection discusses the findings which were revealed by the spatial distributions of Edge Betweenness Centrality (EBC) as well as origin and destination locations. The results were presented in Subsection 5.3.1.

Figure 5.21 has revealed that the demand for taxis is highest in the city centres. It has also revealed that most of the destinations lie within the centres which is why the majority of routes is conducted within these areas. Figure 5.20 has shown that in San Francisco and Vienna, many routes start or end at the highway connecting city centre and airport. However, this is not true as these routes were actually conducted between the airport and the centre but were truncated during the construction of the routes from the FCD.[43]

[43] This issue was already addressed in Subsection 6.1.1 and will therefore not be discussed further here.

With regard to the question of the connection between EBC and the popularity of street segments amongst taxi drivers, the results presented in Subsection 5.3.1 have brought two new insights: firstly, Table 5.3 has shown that the EBC and the popularity of a street segment amongst taxi drivers only correlate weakly in San Francisco and Shanghai but show strong positive correlation in Vienna. Secondly, Figure 5.20 has revealed that there is only weak spatial correlation between street segments with high BC and the street segments most used by taxi drivers because while the drivers mostly operate within the city centre, edges with high BC can be found in all parts of the cities. The results presented in Subsection 5.3.1 cannot fully explain why correlation between EBC and the popularity of a street segment is so high in Vienna because spatial correlation was only accessed visually but not computationally. One possible explanation for the correlation results is that the spatial correlation between EBC and street segments' popularity amongst taxi drivers is higher in Vienna than in San Francisco and Shanghai, which would mean that taxi drivers in San Francisco and Shanghai do not focus on routes with high EBC while in Vienna, the exact opposite is the case. Another possibility is that the spatial correlation is rather low in all three cities, but the correlation between an edge's BC and its popularity is much higher in Vienna which would mean that taxi drivers in none of the three cities focus on street segments with high BC. This would in turn suggest that there are a few very important streets in Vienna which are travelled by a large proportion of the city's taxi trips.
A last point which needs to be addressed when discussing EBC is the use of the normalised global EBC in this thesis (see Subsubsection 2.1.3.2 and Section 4.3). Since the street networks of San Francisco, Shanghai, and Vienna are very different in size, structure, and complexity (see Section 3.1), the EBC values were normalised so that the results could be compared. Normalising EBC is a common approach and has no significant implications (Barthélemy 2018). However, calculating global BC in fragmented networks can lead to significant border effects namely a tendency to compute lower values towards the network's boundaries (Graser et al. 2016). Since the San Francisco street network is a lot smaller than the other two networks, it is particularly affected by this effect. However, as taxi drivers are mostly active in the city centre, the influence of these border effects on the overall results should be marginal.

The analysis of EBC and the locations of origins and destinations has revealed the following findings:

General findings

- Although no final answer could be found regarding spatial correlation between EBC and street segments' popularity amongst taxi drivers, visual assessment did not indicate

spatial correlation between the two.

- Taxi drivers mostly operate in the city centre where demand is high. The only major exception are trips to and from the airport in San Francisco and Vienna.

6.2.2 Road type composition

This subsection discusses the results from the analysis of road types included in taxi drivers' routes. The results were presented in Subsection 5.3.2.

The results have revealed the overall pattern that routes proportionally include less minor roads but more major roads and dual carriageways than the street networks. Since this applies to all route types, it is assumed that the discrepancy is not caused by taxi drivers' route choice behaviour but by the underlying street networks. Figure 5.24 visualises the spatial distribution of the different road types, showing that areas in and around the city centres tend to include proportionally more major roads and less minor roads than the peripheral areas. It also presents the spatial distribution of taxi drivers' actual routes, revealing that they are mostly located in the city centres. This spatial correlation between areas with a high share of major roads and the street segments most used by taxi drivers explains the relatively high proportion of major roads in actual routes.
Figure 5.22 shows that actual and shortest routes present almost identical proportions of dual carriageways whereby the values are much lower than for fastest and fewest intersections routes. This supports the findings from Subsection 6.1.1, stating that taxi drivers tend to prefer optimising for distance instead of time and number of intersections. It seems like drivers in all cities rather take a short route instead of taking a detour over a high performance road although the chosen route is slower and includes more intersections.
Subsection 5.3.2 has revealed that the proportions of certain road types significantly correlate with the Origin-Destination Distance (ODD), suggesting that the longer a taxi ride is, the less minor roads and the more high performance roads it contains. However, the presented results do not explain whether this trend reflects taxi drivers' route choice behaviour or street network characteristics. Calculating the correlations between the road type compositions of actual and optimal routes has revealed that the shares of minor roads, major roads, and dual carriageways show strong and significant positive correlation for all route types in all cities[44], which in turn indicates that the detected patterns are not a consequence of taxi drivers' route choices. Thus, long taxi trips do

[44] Correlation was assessed using the Spearman rank correlation coefficient ρ (Spearman 1904) whereby all results were $\geq$0.48 with p-value $<$0.0001. The results were checked visually with the same kind of density plots presented in Figure 5.23. However, the plots were not included in the thesis as they did not reveal any new information.

not tend to follow a highway because the taxi drivers prefer to do so, but because there are no reasonable alternatives. Apart from the results just discussed, Figure 5.23 has also suggested that in San Francisco, the relationship between road type composition and ODD is reversed when routes become very long, meaning that the proportion of dual carriageways decreases while the proportions of minor and major roads increases. However, this trend is not further discussed because on the one hand the effect is only weak (see Table 5.4) and on the other hand there is almost no data of such long routes available which makes analysis very uncertain (see Subsection 5.3.2).

Another result which needs to be addressed is that in Shanghai, taxi drivers seem to use dual carriageways even if their trip is very short, which is strange because a trip needs to have a certain minimum length in order to include motorway entrance and exit as well as the distance travelled in between.[45] However, the density distribution of the proportion of dual carriageways in short routes presented in Figure 6.5 reveals no odd patterns as routes shorter than 2 kilometres include only very low shares of dual carriageways. Furthermore, only 1.2 % of the routes shorter than 2 kilometres include over 50 % dual carriageways. These results give reason to assume that the relatively high proportion of dual carriageways in short routes in Shanghai is not caused by a data processing error but that it reflects the underlying street network.[46] This hypothesis is also supported by the fact that the Shanghai network proportionally contains more dual carriageways than the two other cities.

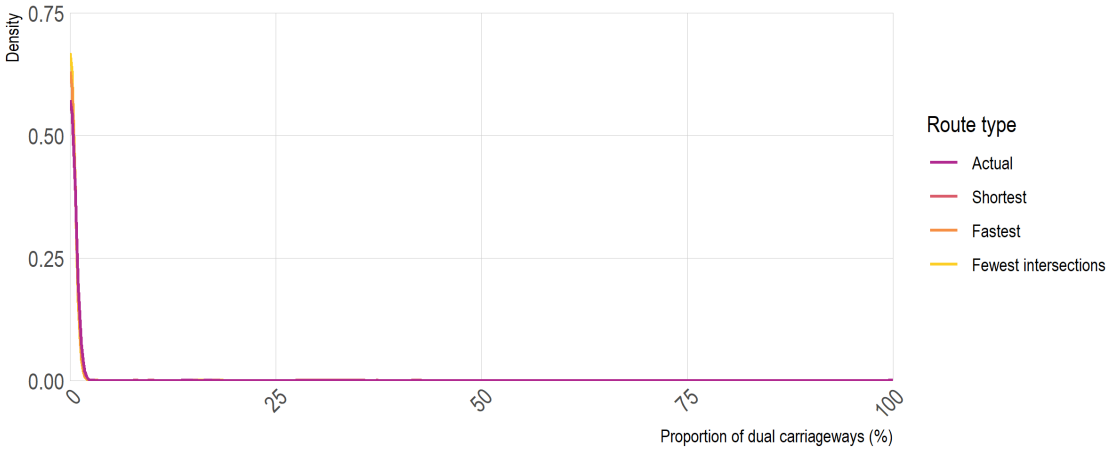


Figure 6.5: Density distribution of proportion of dual carriageways in routes shorter than 2 kilometres in Shanghai.

[45] A potential flaw in the construction of actual routes from the Shanghai FCD is indicated by the city's high proportion of routes with 0 turns which is discussed in Subsection 6.2.4.

[46] Differences between the cities might also originate from the underlying OSM data as the analysis of route type composition is based on OSM road type tags.

The analysis of road type compositions has revealed the following findings:

General findings

- Areas in and near the city centres have proportionally more major roads and less minor roads than peripheral areas.
- Most taxi trips are conducted within the city centre.
- Taxi trips tend to include proportionally more major roads than the respective street networks' average would suggest.
- Taxi drivers tend to avoid detours even if they are faster and have fewer intersections than their chosen route.
- Longer taxi trips are more likely to include proportionally more dual carriageways and less minor roads.

Findings regarding a specific city

- Shanghai is the only city where even routes as short as 2 kilometres include dual carriageways.
- The street network in Shanghai is different from the other two cities' as it includes proportionally more major roads and dual carriageways but less minor roads.

6.2.3 Intersection density and complexity

This subsection discussed the results from the analysis of intersections crossed by taxi drivers' routes. The results were presented in Subsection 5.3.3.

Figure 5.5 has shown that in all three cities, actual routes visit more intersections than optimal routes. This is surprising because taxi drivers have to expect at every intersection that their journey will be delayed by traffic signals and congestion and should therefore be interested in avoiding intersections as much as possible. However, when comparing by the number of intersections a route visits per kilometre instead of the total number of intersections (see Figure 5.25), actual routes show similar or even lower intersection densities than the shortest and fastest alternatives, which in turn suggests that taxi drivers try to avoid intersections.
The results presented in Subsection 5.3.3 have also revealed differences between the cities in terms of routes' and street networks' intersection densities. However, the results of the different route types' intersection densities are similar within each separate city, which indicates that the differences are not caused by the taxi drivers' route choice behaviour but by the underlying street networks. An apparent explanation for the differences would be that any route's intersection density is mainly determined by the street network

and that the routes therefore just reflect their underlying network's intersection density. However, this hypothesis is not supported by Figure 5.25 showing that 1) the intersections densities of the cities' street networks are more similar than their routes' intersection densities and 2) only Shanghai presents similar values for routes and street network while routes present higher intersection densities in San Francisco and Vienna. The differences in San Francisco and Vienna could be explained if the routes were clustered in areas with an above-average number of intersections. However, a comparison of figures 5.20, 5.21, and 5.27 reveals that this is only the case in Vienna but not in San Francisco. Furthermore, the figures show that the spatial correlation between the locations visited by actual routes and the locations of intersections is largest in Shanghai. Thus, if the intersection densities of the routes were primarily determined by the road network, the routes and the network in Shanghai would not have such similar values as the routes are spatially concentrated in a district with an above average intersection density. At this point, the reason for the discrepancies between routes and underlying street networks in terms of number of intersections per kilometre remains unclear.
Figure 5.26 has shown that while all route types show a higher proportion of complex intersections than the respective street network, there are only minor differences between the results of the individual route types. Since not only actual but also optimal routes contain an above-average proportion of complex intersections, this is likely not reflecting taxi drivers' route choice behaviour. In Shanghai and Vienna, the difference can be explained by spatial correlation between the locations visited by the routes and the locations of complex intersections (see figures 5.20, 5.21, and 5.27), meaning that the routes are clustered in the city centre where the density of complex intersections is high. However, this relationship is less clear in San Francisco where the clustering of complex intersections is less extreme.

The analysis of intersection density and complexity has revealed the following findings:

General findings

- The intersection density of a route is not simply determined by the street network.
- Spatial correlation between the locations visited by taxi drivers and the locations of complex intersections cannot fully explain differences in routes' intersection densities between the three cities.

6.2.4 Number of turns and turn characteristics

This subsection discusses the results from the analysis of turns in taxi drivers' routes. The results were presented in Subsection 5.3.4.

The results have only revealed minor differences between the cities in terms of total number of turns per route. This indicates that varying results of the different route types reflect taxi drivers' route choice behaviour rather than differences in the underlying street networks.

Figure 5.28 has shown that the actual routes in all three cities contain similar numbers of intersections as the shortest and fastest alternatives, which is surprising as actual routes are longer and include more intersections than optimal routes (see Figure 5.5). It seems that either the differences in length and number of intersections are not large enough to produce major differences in the routes' numbers of turns or that taxi drivers try to avoid turns and their routes therefore include a similar number of turns as the shortest or fastest alternative despite being longer. The second hypothesis is supported by Figure 5.29, which shows that taxi drivers in Shanghai and Vienna turns at relatively few intersections. A potential explanation why Viennese taxi drivers avoid turns more consequently than their colleagues in San Francisco and Shanghai might be the fact that by the time of data collection, traffic rules in Austria did not allow to turn right on red[47] (Wolfermann, Friedrich, and Fellendorf 2019), while it was allowed by Californian (Newson, Kim, and Gordon 2020) and Chinese law (Tang et al. 2019), which would make right turns in Vienna more costly in comparison to the other two cities. However, Figure 5.30 shows that the relative occurrence of right and left turns as well as of flat and sharp turns is about equal in all cities and for all route types which suggests that taxi drivers do neither avoid nor prefer certain types of turns. This finding contradicts the above-mentioned hypothesis that the "right turn on red"-rule influences taxi drivers' route choice behaviour because if that were the case, drivers in San Francisco and Shanghai would present a preference of right turns. Furthermore, this suggests that taxi drivers do neither avoid nor prefer certain types of turns.

It is worth mentioning that Figure 5.29 indicates a relatively high proportion of routes without any turns in Shanghai. Calculating the proportion of routes with 0 turns for the three cities reveals that in Shanghai, 9.7 % of all routes do not contain any turns while this proportion is much lower in the other two cities with 4.3 % in San Francisco and only 1.7 % in Vienna. Figure 6.6 visualises the spatial distribution of these routes showing that there is almost no variation between the different route types, which indicates that

[47] The "right turn on red"-rule allows drivers to make a right turn against a red traffic light provided they do not interfere with pedestrians, cyclists, or other traffic participants (Newson, Kim, and Gordon 2020).

the reason is not the computation of the routes itself but rather the spatial distribution of origin and destination locations. As most of the streets presenting a high proportion of routes without turns are highways or other high performance streets, it seems that the origin and destination locations of at least some of these routes were not extracted correctly from the FCD (see Subsection 4.4.3) because it is unlikely that passengers are picked up and dropped off on a highway.

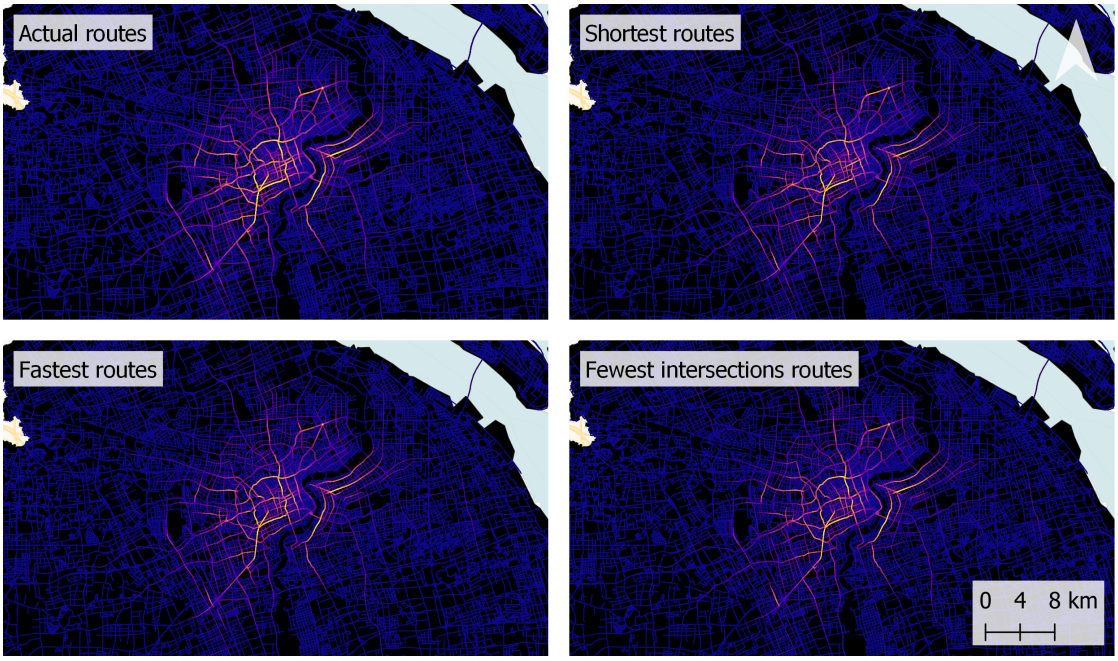


Figure 6.6: Spatial distribution of routes with 0 turns for each route type in Shanghai.

The analysis of turns in taxi drivers' routes has revealed the following findings:

General findings

- The "right turn on red"-rule does not seem to influence taxi drivers' route choice behaviour.

Findings regarding a specific city

- In Shanghai, roughly 10 % of all routes do not contain any turns which is a much larger proportion than in the other two cities. This discrepancy is probably caused by incorrect extraction of origin and destination locations from the Shanghai FCD.

6.3 Synthesis

This section summarises the most relevant findings by revisiting and answering each RQs individually:

RQ1 *How do taxi drivers' routes differ from shortest, fastest, and fewest intersections routes?*

In general, taxi drivers' routes are longer, slower, and include more intersections than the shortest, fastest, and fewest intersections routes. However, the differences are small. On a typical trip, the route chosen by the taxi driver is only about 100 metres longer and takes less than a minute longer than the shortest and fastest route, respectively. Furthermore, it only includes 1 more intersection than the fewest intersections route.

RQ1.1 *Do taxi drivers with passengers on board take the shortest, fastest, or fewest intersections routes?*

Neither the taxi drivers in San Francisco, nor in Shanghai, or in Vienna consistently follow the shortest, fastest, or fewest intersections route. Although, the drivers in all three cities tend to choose routes which are similar to the optimal routes, the extent to which they actually follow them varies between the cities and between individual trips. For taxi drivers in all three cities, length seems to be the most important of the three route selection criteria whereby this is particularly the case when trips are long.

RQ1.2 *How much longer is the actual route than the shortest route, how much slower is the actual route than the fastest route, and how many more intersections does the actual route include than the fewest intersections route?*

Overall, taxi drivers' routes are about 10 % longer than the shortest route and 10–20 % slower than the fastest route. Furthermore, they include about 9 % more intersections than the fewest intersections route.

RQ2 *How does the street network impact taxi drivers' route choice behaviour?*

Taxi drivers operate primarily in the city centre where the density and complexity of the street network is high. Thus, taxi drivers have a large number of alternative routes to choose from and are therefore likely not to follow an optimal route. In contrast, they are likely to follow an optimal route when they travel between the city centre and the airport since there are barely any reasonable alternative routes on this journey. These behavioural patterns show that the drivers are more likely

to choose an optimal route if their options are limited which is particularly the case when the street network is sparse.

RQ2.1 *Is there significant correlation between edge betweenness centrality and the routes chosen by taxi drivers with passengers on board?*

There is significant positive correlation between edge betweenness centrality and the popularity of street segments amongst taxi drivers in all three cities. The correlation is only weak in San Francisco and Shanghai but strong in Vienna.

RQ2.2 *Do taxi drivers with passengers on board avoid or prefer particular road types?*

Taxi drivers with passengers on board do neither avoid nor prefer particular road types.

RQ2.3 *Do taxi drivers with passengers on board avoid or prefer complex intersections?*

The results indicate that taxi drivers neither avoid nor prefer complex intersections. However, a conclusive answer to this RQ requires further research.

RQ2.4 *Do taxi drivers with passengers on board avoid or prefer right or left turns?*

Taxi drivers with passengers on board do neither prefer nor avoid right, left, flat, or sharp turns. However, taxi drivers in all three cities, but particularly in Vienna, try to avoid making turns in general.

RQ3 *How do the findings from RQs 1 and 2 differ among San Francisco, Shanghai, and Vienna?*

In general, the three cities show similar overall trends, which differ in the strength of their characteristics. Regarding the extent to which taxi drivers' routes differ from optimal routes, the results show that the routes chosen by taxi drivers in Shanghai differ the least from optimal routes while the routes chosen by drivers in San Francisco differ the most from optimal routes. Furthermore, the drivers' route choice behaviour differs when trips are long: Taxi drivers in Shanghai and Vienna attach more importance to the length of their route when their trip is long while drivers from San Francisco do not change their route choice behaviour regardless the length of their trip.
Regarding the impact of the street network on taxi drivers' route choice behaviour, no apparent differences between the cities were revealed.

Chapter 7

Conclusions, limitations, and future work

The goal of this thesis was to address the research gap regarding the FCD-based comparison of route choice in different cities as well as to contribute towards a more comprehensive understanding of taxi drivers' route choice in general. For this purpose three large datasets from San Francisco, Shanghai, and Vienna were pre-processed and map matched. Then, the routes chosen by the taxi drivers were reconstructed and a variety of measures was calculated based on three optimal routes, which were computed for each individual trip. Finally, the resulting dataset, containing over 8.5 million routes, was empirically analysed. This novel approach of combining and analysing multiple large-scale datasets has produced the following key findings: Taxi drivers primarily operate in the city centres where the dense and complex street network provides them with many possible routes for their trips. When choosing their routes, the drivers do not simply follow shortest, fastest, or fewest intersections routes but try to avoid turns and choose an alternative which is only partially congruent with the considered optimal routes. For taxi drivers, length seems to be the most important of the three route selection criteria investigated in this thesis, which is a remarkable finding as it differs from previous findings. When choosing their route, taxi drivers tend to avoid detours via the highway, even if this might save time. Drivers in San Francisco behave differently than their colleagues in Shanghai and Vienna in that they do not change their route choice behaviour no matter how long their trip is.

The analysis was based on a selection of route characteristics since including all attributes provided by the routes dataset would certainly have gone beyond the scope of this thesis. Thus, the potential findings the data might further reveal are manifold as the dataset includes multiple additional route attributes which have not yet been investigated. This

thesis therefore not only provides new knowledge about the route choice behaviour of taxi drivers but also a basis for further research in this field.

The analysis presented in this thesis has shown some limitations which are partly due to the chosen methodology and partly inherent to the chosen approach and data. The most significant limitation in terms of methodology is arguably the route construction procedure, which has led to the truncation of long trips in San Francisco and, to a lesser extent, Vienna. Another limitation is the exclusion of contextual information such as traffic, whether, or information about the use of navigational aids.
The third aspect which needs to be mentioned here are the implications of using FCD and OSM data. The three sets of FCD were collected in 2008, 2010, and 2015 while the street network data represents the situation in 2019. Structural changes to the street networks could therefore mean that the routes and the network do not match in some places. Regarding the FCD, it is inevitable that the results of FCD-based analyses are subject to a certain degree of uncertainty, since FCD always contain positional inaccuracies which cannot be completely eliminated by map matching. However, this is not considered a limitation in this work as large datasets were used to minimise error influences and a lot of attention was paid to map matching in order to further lower data-related uncertainty. Finally, it should be pointed out that the use of OSM data always involves some uncertainty since OSM is a crowd-based project. In this thesis, this is particularly important in terms of the analysis of road types as it is likely that some roads are misclassified (Zhang and Malczewski 2017).

Addressing the limitations mentioned above, it is suggested that future work improves the route construction procedure and includes contextual information to get a more comprehensive picture of the manifold drivers of route choice behaviour. The truncation of routes has also revealed that the study area should not be defined by political boundaries but by the taxi drivers' movement patterns. Furthermore, it was shown that for very short or very long routes, there are often no reasonable alternative routes for the drivers to choose from. It is therefore suggested that not only the similarity between actual and optimal routes, but also the similarity between the different alternatives, is assessed in order to estimate how large the taxi drivers' choice set is. For the assessment of route similarity, a dual approach based on PSL and FD is recommended since the strength of one measure corresponds to the weakness of the other. The results in this thesis have further indicated that considering the street network on a city level is not sufficient when accessing the network's influence on drivers' route choices, since they are bound to their immediate surroundings when choosing their routes. Future studies should therefore consider the network on a much smaller scale.

Since it has become clear that taxi drivers do not follow shortest, fastest, or fewest intersections routes, the range of potential route characteristics determining drivers' routes should be extended. In this context, the finding that taxi drivers tend to avoid making turns seems to be a promising start.

Bibliography

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). *Network Flows: Theory, Algorithms, and Applications*. Upper Saddle River: Prentice Hall.

Alt, H. and Guibas, L. J. (2000). "Discrete Geometric Shapes: Matching, Interpolation, and Approximation". In: *Handbook of Computational Geometry*. Ed. by J.-R. Sack and J. Urrutia. Amsterdam: Elsevier. Chap. 3, pp. 121–153.

Barabási, A.-L. (2002). *Linked: The new science of networks*. Cambridge: Perseus.

Barnes, J. A. and Harary, F. (1983). "Graph theory in network analysis". *Social Networks* 5(2), pp. 235–244.

Barthélemy, M. (2011). "Spatial networks". *Physics Report* 499(1–3), pp. 1–101.

Barthélemy, M. (2014). "Spatial Networks". In: *Encyclopedia of Social Network Analysis and Mining*. Ed. by R. Alhajj and J. Rokne. New York: Springer, pp. 1967–1976.

Barthélemy, M. (2018). *Morphogenesis of spatial networks*. Berlin: Springer.

Barua, S. (2019). "A Discrete Route Choice Model Using Support Vector Machine an Context of Dhaka City". *Daffodil International University Journal of Science and Technology* 14(1), pp. 28–31.

Bayer, M. (2012). "SQLAlchemy". In: *The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks*. Ed. by A. Brown and G. Wilson. Mountain View: aosabook.org. Chap. 20.

Beijing Expat Service Center (2019). *Speed Limits in China*. https://www.beijingesc.com/tips-on-driving-in-china/20-on-speed-limits.html. last visited on April 10, 2020.

Bellman, R. (1958). "On a routing problem". *Quarterly of Applied Mathematics* 16(1), pp. 87–90.

Ben-Akiva, M. E. (1973). "Structure of passenger travel demand models". Ph.D. Thesis. Cambridge: Massachusetts Institute of Technology.

Ben-Akiva, M. E. (1974). "Structure of passenger travel demand models". *Transportation Research Record* 526(1), pp. 26–42.

Ben-Akiva, M. E., Bergman, M. J., Daly, A. J., and Ramaswamy, R. (1984). "Modeling inter-urban route choice behaviour". In: *Proceedings of the Ninth International Symposium on Transportation and Traffic Theory*. Ed. by J. Volmuller and R. Hamerslag. VNU Science Press, pp. 299–330.

Ben-Akiva, M. E. and Bierlaire, M. (1999). "Discrete choice methods and their applications to short term travel decisions". In: *Handbook of Transportation Science*. Ed. by R. W. Hall. Dordrecht: Kluwer Academic Publishers. Chap. 2, pp. 5–33.

Ben-Akiva, M. E. and Bierlaire, M. (2003). "Discrete Choice Models with Applications to Departure Time and Route Choice". In: *Handbook of Transportation Science*. Ed. by R. W. Hall. Boston: Springer. Chap. 2, pp. 7–37.

Ben-Akiva, M. E., Ramming, M. S., and Bekhor, S. (2004). "Route Choice Models". In: *Human Behaviour and Traffic Networks*. Ed. by M. Schreckenberg and R. Selten. Berlin, Heidelberg: Springer, pp. 23–45.

Bhatia, S. (2019). "Survey of shortest Path Algorithms". *International Journal of Computer Science and Engineering* 6(11), pp. 33–39.

Boeing, G. (2017). "OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks". *Computers, Environment and Urban Systems* 65(1), pp. 126–139.

Bouton, S., Knupfer, S. M., Mihov, I., and Swartz, S. (2015). *Urban mobility at a tipping point*. https://www.mckinsey.com/business-functions/sustainability/our-insights/urban-mobility-at-a-tipping-point. last visited on September 28, 2019.

Bovy, P. H. and Stern, E. (1990). *Route Choice: Wayfinding in Transport Networks*. Dordrecht: Kluwer Academic Publishers.

Boxer, B. (2019). *Shanghai*. https://www.britannica.com/place/Shanghai. last visited on March 27, 2020.

Brakatsoulas, S., Pfoser, D., Salas, R., and Wenk, C. (2005). "On Map-Matching Vehicle Tracking Data". In: *Proceedings of the 31st international conference on Very large data bases (VLDB '05)*. VLDB Endowmnent, pp. 853–864.

Brandes, U. (2001). "A faster algorithm for betweenness centrality". *The Journal of Mathematical Sociology* 25(2), pp. 163–177.

Buchanan, M. (2002). *Nexus: Small Worlds and the Groundbreaking Science of Networks*. New York: Norton.

Cai, H., Zhan, X., Zhu, J., Jia, X., Chiu, A. S. F., and Xu, M. (2016). "Understanding taxi travel patterns". *Physica A: Statistical Mechanics and its Applications* 457(1), pp. 590–597.

Cascetta, E., Nuzzolo, A., Russo, F., and Vitetta, A. (1996). "A modified logit route choice model overcoming path overlapping problems. Specification and some calibration results for interurban networks". In: *Transportation and Traffic Theory. Proceedings of The 13th International Symposium On Transportation And Traffic Theory*. Ed. by J. B. Lesort. Pergamon, pp. 697–711.

Chen, F., Shen, M., and Tang, Y. (2011). "Local Path Searching Based Map Matching Algorithm for Floating Car Data". *Procedia Environmental Sciences* 10(A), pp. 576–582.

Chiu, Y.-C., Bottom, J., Mahut, M., Paz, A., Balakrishna, R., Waller, T., and Hicks, J. (2011). *Dynamic Traffic Assignment: A Primer*. Washington D.C.: Transportation Research E-Circular.

Chorley, R. J. and Haggett, P., eds. (1967). *Models in Geography*. London: Methuen and Co.

Chu, C. (1989). "A paired combinatorial logit model for travel demand analysis". In: *Proceedings of the Fifth World Conference on Transportation Research*, pp. 295–309.

Cintula, P., Fermüller, C. G., and Noguera, C. (2017). "Fuzzy Logic". In: *The Stanford Encyclopedia of Philosophy (Fall 2017 Edition)*. Ed. by E. N. Zalta. https://plato.stanford.edu/archives/fall2017/entries/logic-fuzzy/. last visited on April 25, 2020.

City of Vienna (2012). *Höherrangiges Strassennetz Wien – Bestand und Planung*. https://www.wien.gv.at/stadtentwicklung/projekte/verkehrsplanung/strassen/bundesstrassen/. last visited on June 21, 2020.

City of Vienna (2020). *Zahlen und Fakten zum Wiener Straßennetz*. https://www.wien.gv.at/verkehr/strassen/fakten.html. last visited on March 27, 2020.

Cohen, J. (1992). "A power primer". *Psychological Bulletin* 122(1), pp. 155–159.

COMSIS Corporation (1995). *Analysis of travelers' preferences for routing: literature review*. Report submitted to Federal Highway Administration, U.S. Department of Transportation, DTFH61-95-C-00017.

Conrad, B., Hansen, G. C., Lamott, K., and The Editors of Encyclopaedia Britannica (2019). *San Francisco*. https://www.britannica.com/place/San-Francisco-California. last visited on March 26, 2020.

Correa, J. R. and Stier-Moses, N. E. (2011). "Wardrop Equilibria". In: *Wiley Encyclopedia of Operations Research and Management Science*. Ed. by J. J. Cochran, L. A. Cox, P. Keskinocak, J. P. Kharoufeh, and J. C. Smith. Hoboken: John Wiley & Sons.

Crucitti, P., Latora, V., and Porta, S. (2006). "Centrality in networks of urban streets". *Chaos: An Interdisciplinary Journal of Nonlinear Science* 16(1), Article no. 015113.

Cullen, A. C. and Frey, H. C. (1999). *Probabilistic Techniques in Exposure Assessment*. New York: Plenum Press.

D'Andrea, E. and Marcelloni, F. (2017). "Detection of traffic congestion and incidents from GPS trace analysis". *Expert Systems with Applications* 73(1), pp. 43–56.

Daganzo, C. F. and Sheffi, Y. (1977). "On Stochastic Models of Traffic Assignment". *Transportation Science* 11(3), pp. 253–274.

Dargay, J., Gately, D., and Sommer, M. (2007). "Vehicle Ownership and Income Growth, Worldwide: 1960–2030". *The Energy Journal* 28(4), pp. 143–170.

De La Barra, T., Pérez, B., and Añez, J. (1993). "Multidimensional path search and assignment". In: *Proceedings of the 21st PTRC Summer Annual Meeting*. University of Manchester, pp. 307–319.

Dhulipala, S., Kedia, A. S., Salini, P. S., and Katti, B. K. (2017). "Building A Neuro-Fuzzy Based Route Choice Model in Metropolitan Context: Surat City in India". *Transportation Research Procedia* 25(1), pp. 3203–3219.

Diestel, R. (2017). "Extremal Graph Theory". In: *Graph Theory*. Ed. by S. Axler and K. Ribet. Graduate Texts in Mathematics. Berlin: Springer. Chap. 7, pp. 173–208.

Dijkstra, E. W. (1959). "A Note on Two Problems in Connexion with Graphs". *Numerische Mathematik* 1(1), pp. 269–271.

Domingos, P. (2012). "A few useful things to know about machine learning". *Communication of the ACM* 55(10), pp. 78–87.

Ehrlich, B., Holzner, L., Hill, R. J., and Michael, D. (2020). *Vienna*. https://www.britannica.com/place/Vienna. last visited on March 27, 2020.

Eiter, T. and Mannila, H. (1994). *Computing Discrete Fréchet Distance*. CD-TR 94/64. Vienna: Christian Doppler Laboratory for Expert Systems, TU Vienna.

Euler, L. (1741). "Solutio problematis ad geometriam situs pertinentis". *Commentarii academiae scientiarum Petropolitanae* 8(1), pp. 128–140.

Feld, S. (2019). *Alternative Routen in komplexen Umgebungen*. Wiesbaden: Springer Vieweg.

Florian, M. (1999). "Untangling traffic congestion: application of network equilibrium models in transportation planning". *ORMS Today* 26(2), pp. 52–57.

Floyd, R. W. (1962). "Algorithm 97: shortest path". *Communications of the ACM* 5(6), p. 345.

Ford, L. R. (1956). *Network flow theory, Paper P-923*. Santa Monica: The Rand Corporation.

Fréchet, M. R. (1906). "Sur quelques points du calcul fonctionnel". *Rendiconti del Circolo Matematico di Palermo (1884–1940)* 22(1), pp. 1–72.

Frejinger, E., Bierlaire, M., and Ben-Akiva, M. (2009). "Sampling of alternatives for route choice modeling". *Transportation Research Part B: Methodological* 43(10), pp. 984–994.

Fürnkranz, J. (2011). "Decision Tree". In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Boston: Springer, pp. 263–267.

Gärling, T. (1998). "Behavioural assumptions overlooked in travel choice modelling". *Travel Behaviour Research: Updating the State of Play*, pp. 3–18.

Genolini, C. (2016). *kmlShape: K-Means for Longitudinal Data using Shape-Respecting Distance*. https://CRAN.R-project.org/package=kmlShape. last visited on June 9, 2020.

GeoPandas Development Team (2020). *GeoPandas: Python tools for geographic data*. https://github.com/geopandas/geopandas. last visited on April 4, 2020.

George, B. (2016). "Graph Theory, Konigsberg Problem". In: *Encyclopedia of GIS*. Ed. by S. Shekar, H. Xiong, and X. Zhou. Cham: Springer.

Gillies, S. et al. (2007). *Shapely: manipulation and analysis of geometric objects*. https://github.com/Toblerity/Shapely. last visited on April 4, 2020. toblerity.org.

Graser, A., Leodolter, M., Koller, H., and Brändle, N. (2016). "Improving vehicle speed estimates using street network centrality". *International Journal of Cartography* 2(1), pp. 77–94.

Greenshields, B. D., Bibbins, J. R., Channing, W. S., and Miller, H. H. (1935). "A Study of Traffic Capacity". *Highway Research Board* 14(1), pp. 448–477.

Gupta, R. and Pathak, C. (2014). "A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing". *Procedia Computer Science* 36(1), pp. 599–605.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). "Exploring network structure, dynamics, and function using NetworkX". In: *Proceedings of the 7th Python in Science Conference (SciPy2008)*. Ed. by G. Varoquaux, T. Vaught, and J. Millman, pp. 11–15.

Haggett, P. and Chorley, R. J. (1969). *Network analysis in geography*. London: Edward Arnold.

Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). "A formal basis for the heuristic determination of minimum cost paths". *IEEE transactions on Systems Science and Cybernetics* 4(2), pp. 100–107.

Hausdorff, F. (1927). *Mengenlehre*. Berlin: De Gruyter & Co.

Hawas, Y. E. (2002). "Developing fuzzy route choice models using neural nets". In: *Intelligent Vehicle Symposium*. IEEE, pp. 71–76.

Henn, V. (2003). "Route Choice Making Under Uncertainty: a Fuzzy Logic Based Approach". In: *Fuzzy Sets Based Heuristics for Optimization*. Ed. by J.-L. Verdegay. Studies in Fuzziness and Soft Computing, vol 126. Berlin: Springer, pp. 277–292.

Hillier, B. and Hanson, J. (1984). *The Social Logic of Space*. Cambridge: Cambridge University Press.

Hopkins, B. and Wilson, R. J. (2004). "The Truth about Königsberg". *The College Mathematics Journal* 35(3), pp. 198–207.

Hunter, T., Abbeel, P., and Bayen, A. (2014). "The path inference filter: model-based low-latency map matching of probe vehicle data". *IEEE Transactions on Intelligent Transportation Systems* 15(2), pp. 507–529.

Jahangiri, A. and Rakha, H. A. (2015). "Applying Machine Learning Techniques to Transportation Mode Recognition Using Mobile Phone Sensor Data". *IEEE Transactions on Intelligent Transportation Systems* 16(5), pp. 2406–2417.

Jayakrishnan, R., Tsai, W. K., and Chen, A. (1995). "A dynamic traffic assignment model with traffic-flow relationships". *Transportation Research Part C: Emerging Technologies* 3(1), pp. 51–72.

Jensen, C. S. and Tradišauskas, N. (2009). "Map Matching". In: *Encyclopedia of Database Systems*. Ed. by L. Liu and M. T. Özsu. Boston: Springer, pp. 1692–1696.

Jiang, B., Yin, J., and Zhao, S. (2009). "Characterizing the human mobility pattern in a large street network". *Physical Review E* 80(2). Article no. 021136.

Jing, N., Huang, Y.-W., and Rundensteiner Elke, A. (1998). "Hierarchical Encoded Path Views for Path Query Processing: An Optimal Model and Its Performance Evaluation". *IEEE Transactions on Knowledge and Data Engineering* 10(3), pp. 409–432.

Jing, P., Zhao, M., He, M., and Chen, L. (2018). "Travel Mode and Travel Route Choice Behavior Based on Random Regret Minimization: A Systematic Review". *Sustainability* 10(4), Article no. 1185.

Johnson, D. B. (1977). "Efficient algorithms for shortest paths in sparse networks". *Journal of the ACM (JACM)* 24(1), pp. 1–13.

Kim, J. and Mahmassani, H. S. (2015). "Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories". *Transportation Research Procedia* 9(1), pp. 164–184.

Koller, H., Widhalm, P., Dragaschnig, M., and Graser, A. (2015). "Fast Hidden Markov Model Map-Matching for Sparse and Noisy Trajectories". In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC 2015)*. IEEE, pp. 2557–2561.

Koppelman, F. S. and Wen, C.-H. (2000). "The paired combinatorial logit model: properties, estimation and application". *Transportation Research Part B: Methodological* 34(2), pp. 75–89.

Krishna, S., Katti, B. K., and Gaurang, J. (2015). "Literature Review of Traffic Assignment: Static and Dynamic". *International Journal of Transportation Engineering* 2(4), pp. 339–347.

Kruppa, J., Schwarz, A., Arminger, G., and Ziegler, A. (2013). "Consumer credit risk: Individual probability estimates using machine learning". *Expert Systems with Applications* 40(13), pp. 5152–5131.

Kühne, R., Schäfer, R.-P., Mikat, J., Thiessenhusen, K.-U., Böttger, U., and Lorkowski, S. (2003). "New Approaches for Traffic Management in Metropolitan Areas". *IFAC Proceedings Volumes* 36(14), pp. 209–214.

Kung, R.-M., Hanson, E. N., Ioannidis, Y. E., Sellis, T. K., Shapiro, L. D., and Stonebraker, M. (1984). "Heuristic search in database systems". *Expert Database Workshop*, pp. 537–548.

Lai, X., Fu, H., Li, J., and Sha, Z. (2019). "Understanding drivers' route choice behaviours in the urban network with machine learning models". *IET Intelligent Transport Systems* 13(3), pp. 427–434.

Lämmer, S., Gehlsen, B., and Helbing, D. (2006). "Scaling laws in the spatial structure of urban road networks". *Physica A: Statistical Mechanics and its Applications* 363(1), pp. 89–95.

Li, A. C. Y., Nozick, L., Davidson, R., and Brown, N. (2013). "Approximate Solution Procedure for Dynamic Traffic Assignment". *Journal of Transportation Engineering* 139(8), pp. 822–832.

Li, B., Guo, Y., Zhou, J., and Cai, Y. (2018). "A Data Correction Algorithm for Low-Frequency Floating Car Data". *Sensors* 18(1). Article no. 3639.

Li, L., Wang, S., and Wang, F.-Y. (2018). "An Analysis of Taxi Driver's Route Choice Behavior Using the Trace Records". *IEEE Transactions on Computational Social Systems* 5(2), pp. 576–582.

Lin, T. (2019). *Benchmark of popular graph/network packages*. https://www.timlrx.com/2019/05/05/benchmark-of-popular-graph-network-packages/. last visited on April 6, 2020.

Liu, K. and Xu, Y. (2019). "Route Choice Behavior: Understanding the Impact of Asymmetric Preference on Travelers' Decision Making". *Symmetry* 11(1), Article no. 66.

Liu, L., Andris, C., and Ratti, C. (2010). "Uncovering cabdrivers' behavior patterns from their digital traces". *Computers, Environment and Urban Systems* 34(1), pp. 541–548.

Liu, X., Gong, L., Gong, Y., and Liu, Y. (2015). "Revealing travel patterns and city structure with taxi trip data". *Journal of Transport Geography* 43(1), pp. 78–90.

Lou, Y., Zhang, C., Zheng, Y., Xie, X., Wang, W., and Huang, Y. (2009). "Map-matching for low-sampling-rate GPS trajectories". In: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09)*. ACM, pp. 352–361.

Luce, R. D. (1959). *Individual choice behavior: a theoretical analysis*. New York: Wiley.

Ma, X., Dai, Z., He, Z., Ma, J., Wang, Y., and Wang, Y. (2017). "Learning Traffic as Images: A Deep Convolutional Neural Network for Large-Scale Transportation Network Speed Prediction". *Sensors* 17(4). Article no. 818.

Ma, X., Yu, H., Wang, Y., and Wang, Y. (2015). "Large-Scale Transportation Network Congestion Evolution Prediction Using Deep Learning Theory". *PLoS ONE* 10(3). Article no. e0119044.

Madkour, A., Aref, W. G., Rehman, F. U., Rahman Mohamed, A., and Basalamah, S. (2017). *A Survey of Shortest-Path Algorithms*. arXiv eprint 1705.02044. https://arxiv.org/abs/1705.02044. last visited on April 23, 2020.

Manley, E. J., Orr, S. W., and Cheng, T. (2015). "A heuristic model of bounded route choice in urban areas". *Transportation Research Part C: Emerging Technologies* 56(1), pp. 159–209.

Manley, E., Addison, J. D., and Cheng, T. (2015). "Shortest path or anchor-based route choice: a large-scale empirical analysis of minicab routing in London". *Journal of Transport Geography* 43(1), pp. 123–139.

Mann, H. and Whitney, D. R. (1947). "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". *Annals of Mathematical Statistics* 16(1), pp. 50–60.

Masuda, N., Porter, M. A., and Lambiotte, R. (2017). "Random walks and diffusion on networks". *Physics Reports* 716/717, pp. 1–58.

McFadden, D. (1978). "Modeling the choice of residential location". *Transportation Research Record* 673(1), pp. 72–77.

McKinney, W. (2011). "pandas: a foundational Python library for data analysis and statistics". *Python for High Performance and Scientific Computing* 14.

Meng, Y., Chen, W., Li, Z., Chen, Y., and Chao, J. C. H. (2009). "A Simplified Map-Matching Algorithm for In-Vehicle Navigation Unit". *Geographic Information Sciences* 8(1), pp. 24–30.

Messelodi, S., Modena, C. M., Zanin, M., Granelli, F., De Natale, F. G. B., Betterle, E., and Guarise, A. (2009). "Intelligent extended floating car data collection". *Expert Systems with Applications* 36(1), pp. 4213–4227.

Miwa, T., Kiuchi, D., Yamamoto, T., and Morikawa, T. (2012). "Development of map matching algorithm for low frequency probe data". *Transportation Research Part C: Emerging Technologies* 22(1), pp. 132–145.

Miyagi, T. (2004). "A modelling of route choice behaviour in transportation networks: an approach from reinforcement learning". In: *Urban Transport X. Urban Transport and the Environment in the 21st Century*. Ed. by C. A. Brebbia and L. C. Wadhwa. WIT Press, pp. 235–244.

Moore, E. F. (1959). "The shortest path through a maze". In: *Proceedings of an International Symposium on the Theory of Switching*. Harvard University Press, pp. 285–292.

Morikawa, T., Miwa, T., Kurauchi, S., Yamamoto, T., and Kobayashi, K. (2005). "Driver's Route Choice Behavior and its Implications on Network Simulation and Traffic Assignment". In: *Simulation Approaches in Transportation Analysis: Recent Advances and Challenges*. Operations Research/Computer Science Interfaces Series, vol 31. Boston: Springer, pp. 341–369.

Murat, Y. S. and Uludag, N. (2008). "Route choice modelling in urban transportation networks using fuzzy logic and logistic regression methods". *Journal of Scientific & Industrial Research* 67(1), pp. 19–27.

Nash, J. F. (1950). "Equilibrium points in n-person games". *Proceedings of the National Academy of Sciences USA* 36(1), pp. 48–49.

Nash, J. F. (1951). "Non-cooperative games". *Annals of Mathematics* 54(1), pp. 286–295.

National Research Council (2005). *Network Science*. Washington D.C.: The National Academy Press.

Newman, M., Barabási, A.-L., and Watts, D. J. (2006a). "Introduction". In: *The Structure and Dynamics of Networks*. Ed. by M. Newman, A.-L. Barabási, and D. J. Watts. Princeton and Oxford: Princeton University Press. Chap. 1, pp. 1–8.

Newman, M., Barabási, A.-L., and Watts, D. J., eds. (2006b). *The Structure and Dynamics of Networks*. Princeton and Oxford: Princeton University Press.

Newson, G., Kim, D. S., and Gordon, S. (2020). *California: Driver Handbook*. Sacramento: California State Transportation Agency.

Newson, P. and Krumm, J. (2009). "Hidden Markov map matching through noise and sparseness." In: *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (GIS '09)*. ACM, pp. 336–343.

Nian, G., Zhu, W., and Sun, J. (2017). "Analyzing behavior differences of occupied and non-occupied taxi drivers using floating car data". *Journal of Shanghai Jiaotong University (Science)* 22(6), pp. 682–687.

Ochieng, W. Y., Quddus, M. A., and Noland, R. B. (2003). "Map-matching in complex urban networks". *Brazilian Journal of Cartography* 55(2), pp. 1–18.

Oliphant, T. E. (2006). *A guide to NumPy*. USA: Trelgol Publishing.

OpenStreetMap contributors (2017). *Planet dump retrieved from https://planet.osm.org*. https://www.openstreetmap.org. last visited on March 25, 2020.

OpenStreetMap contributors (2018). *Why OpenStreetMap?* https://wiki.openstreetmap.org/wiki/Why_OpenStreetMap%3F. last visited on March 25, 2020.

OpenStreetMap contributors (2019). *About OpenStreetMap*. https://wiki.openstreetmap.org/wiki/About_OpenStreetMap. last visited on March 25, 2020.

OpenStreetMap contributors (2020a). *California*. https://wiki.openstreetmap.org/wiki/California. last visited on April 10, 2020.

OpenStreetMap contributors (2020b). *Key:highway*. https://wiki.openstreetmap.org/wiki/Key:highway#Roads. last visited on April 10, 2020.

OpenStreetMap contributors (2020c). *OSM tags for routing/Maxspeed*. https://wiki.openstreetmap.org/wiki/OSM_tags_for_routing/Maxspeed. last visited on April 10, 2020.

Osborne, M. J. and Rubinstein, A. (1994). *A Course in Game Theory*. Cambridge, Massachusetts: The MIT Press.

Özkan, T., Lajunen, T., Chliaoutakis, J. E., Parker, D., and Summala, H. (2006). "Cross-cultural differences in driving behaviours: A comparison of six countries". *Transportation Research Part F: Traffic Psychology and Behaviour* 9(3), pp. 227–242.

Pallottino, S. and Scutellà, M. G. (1998). "Shortest Path Algorithms in Transportation Models: Classical and Innovative Aspects". In: *Equilibrium and Advanced Transportation Modelling*. Ed. by P. Marcotte and S. Nguyen. Boston: Springer. Chap. 11, pp. 245–281.

Paulin, T. and Bessler, S. (2013). "Controlled Probing – A system for targeted floating car data collection". In: *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*. IEEE, pp. 1095–1100.

Peixoto, T. P. (2014). *The graph-tool python library*. https://figshare.com/articles/graph_tool/1164194. last visited on April 4, 2020.

Peixoto, T. P. (2020). *graph_tool.topology – Assessing graph topology*. https://graph-tool.skewed.de/static/doc/topology.html. last visited on Mai 11, 2020.

Peixoto, T. P. (2015). *Graph-tool performance comparison*. https://graph-tool.skewed.de/performance. last visited on April 6, 2020.

Peng, C., Jin, X., Wong, K.-C., Shi, M., and Liò, P. (2012). "Collective Human Mobility Pattern from Taxi Trips in Urban Area". *PLoS ONE* 7(8). Article no. e34487.

Pettie, S. (2008). "Single-Source Shortest Paths". In: *Encyclopedia of Algorithms*. Ed. by M.-Y. Kao. Boston: Springer, pp. 847–849.

Pfoser, D. (2008). "Floating Car Data". In: *Encyclopedia of GIS*. Ed. by S. Shekar and H. Xiong. Cham: Springer, p. 321.

Pfoser, D., Brakatsoulas, S., Brosch, P., Umlauft, M., Tryfona, N., and Tsironis, G. (2008). "Dynamic travel time provision for road networks". In: *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. Article no. 68. ACM.

Pfoser, D. and Jensen, C. S. (1999). "Capturing the Uncertainty of Moving-Object Representations". In: *Advances in Spatial Databases. SSD 1999. Lecture Notes in Computer Science, vol 1651*. Ed. by R. H. Güting, D. Papadias, and F. Lochovsky. Berlin, Heidelberg: Springer, pp. 111–131.

Piórkowski, M., Sarafijanovoc-Djukic, N., and Grossglauser, M. (2009). "A Parsimonious Model of Mobile Partitioned Networks with Clustering". In: *The First International Conference on COMmunication Systems and NETworkS (COMSNETS)*. IEEE, pp. 1–10.

Porta, S., Crucitti, P., and Latora, V. (2006a). "The network analysis of urban streets: a dual approach". *Physica A: Statistical Mechanics and its Applications* 369(2), pp. 853–866.

Porta, S., Crucitti, P., and Latora, V. (2006b). "The network analysis of urban streets: a primal approach". *Environment and Planning B: Planning and Design* 33(5), pp. 705–725.

Prashker, J. N. and Bekhor, S. (2004). "Route Choice Models in the Stochastic User Equilibrium Problem: A Review". *Transport Reviews* 24(4), pp. 437–463.

Prato, C. G. (2009). "Route choice modeling: past, present and future research directions". *Journal of Choice Modelling* 2003(1), pp. 64–73.

QGIS Development Team (2020). *QGIS Geographic Information System*. Open Source Geospatial Foundation Project. http://qgis.osgeo.org. last visited on April 4, 2020.

Quddus, M. A., Noland, R. B., and Ochieng, W. Y. (2006). "A High Accuracy Fuzzy Logic Based Map Matching Algorithm for Road Transport". *Journal of Intelligent Transportation Systems* 10(3), pp. 103–115.

Quddus, M. A., Ochieng, W. Y., and Noland, R. B. (2007). "Current map-matching algorithms for transport applications: State-of-the art and future research directions". *Transportation Research Part C: Emerging Technologies* 15(5), pp. 312–328.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/. last visited on April 4, 2020.

Rabiner, L. and Juang, B. (1986). "An introduction to hidden Markov models". *IEEE ASSP Magazine* 32(1), pp. 4–16.

Ramming, M. S. (2002). "Network knowledge and route choice". Ph.D. Thesis. Cambridge: Massachusetts Institute of Technology.

Rey, S. J. (2015). "Mathematical Models in Geography". In: *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*. Ed. by J. D. Wright. Amsterdam: Elsevier, pp. 785–790.

Rodrigue, J.-P., Comtois, C., and Slack, B. (2017). *The Geography of Transport Systems*. 4th ed. https://transportgeography.org/. last visited on Oct 30, 2019. New York: Routledge.

Ruphail, N. M., Ranjithan, S. R., El Dessouki, W., Smith, T., and Brill, E. D. (1996). "A Decision Support System for Dynamic Pre-Trip Route Planning". In: *Applications of Advanced Technologies in Transportation Engineering: Proceedings of The Fourth International Conference*, pp. 325–329.

Rust, J. (2016). "Dynamic Programming". In: *The New Palgrave Dictionary of Economics*. Ed. by M. Vernengo, E. P. Caldentey, and B. J. Rosser Jr. London: Palgrave Macmillan.

Sammut, C. (2011). "Greedy Search". In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Boston: Springer, pp. 482–483.

Sammut, C. and Webb, G. I. (2011). "Preface". In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Boston: Springer.

Schäfer, R.-P., Thiessenhusen, K.-U., and Wagner, P. (2002). *A traffic information system by means of real-time floating-car data*. Conference Paper. ITS World Congress 2002.

Sheffi, Y. and Powell, W. B. (1982). "An algorithm for the equiliibrium assignment problem with random link times". *Networks* 12(2), pp. 191–207.

Sheffi, Y. (1985). *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*. Englewood: Prentice-Hall.

Sniedovich, M. (2006). "Dijkstra's algorithm revisited: the dynamic programming connexion". *Control and Cybernetics* 35(3), pp. 599–620.

Spearman, C. (1904). "The Proof and Measurement of Association between Two Things". *American Journal of Psychology* 15(1), pp. 72–101.

State of California (2011). *California County Population Estimates and Components of Change by Year, July 1, 2000-2010*. Sacramento: Department of Finance.

Statistical Office of the City of Vienna (2015). *Statistisches Jahrbuch der Stadt Wien*. Vienna: Magistrat der Stadt Wien.

Sun, B. and Park, B. B. (2017). "Route choice modeling with Support Vector Machine". *Transportation Research Procedia* 25(1), pp. 1806–1814.

Sun, D. J., Zhang, C., Zhang, L., Chen, F., and Peng, Z.-R. (2014). "Urban travel behavior analyses and route prediction based on floating car data". *Transportation Letters* 6(3), pp. 118–125.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. Cambridge: MIT Press.

Tang, J., Liu, F., Wang, Y., and Wang, H. (2015). "Uncovering urban human mobility from large scale taxi GPS data". *Physica A: Statistical Mechanics and its Applications* 438(1), pp. 140–153.

Tang, J., Liu, F., Zou, Y., Zhang, W., and Wang, Y. (2017). "An Improved Fuzzy Neural Network for Traffic Speed Prediction Considering Periodic Characteristic". *IEEE Transactions on Intelligent Transportation Systems* 18(9), pp. 2340–2350.

Tang, K., Li, K., Li, M., and Liu, D. (2019). "China". In: *Global Practices on Road Traffic Signal Control*. Ed. by K. Tang, M. Boltze, H. Nakamura, and Z. Tian. Amsterdam: Elsevier. Chap. 11, pp. 185–215.

Taylor, H. A., Gardony, A. L., and Brunyé, T. T. (2018). "Environmental knowledge: cognitive flexibility in structures and process". In: *Handbook of Behavioral and Cognitive Geography*. Ed. by D. R. Montello. Cheltenham: Edward Elgar. Chap. 6, pp. 97–115.

The PostgreSQL Global Development Group (2020). *PostgreSQL 10.12 Documentation*. https://www.postgresql.org/docs/10/index.html. last visited on April 4, 2020.

United Nations (2018). *The World's Cities in 2018 – Data Booklet*. Department of Economic and Social Affairs, Population Division. Document ST/ESA/SER.A/417.

United Nations (2019). *World Population Prospects 2019: Highlights*. Department of Economic and Social Affairs, Population Division. Document ST/ESA/SER.A/423.

Van Diggelen, T. W. T. (2018). "Efficiently answering Fréchet queries". MA thesis. Eindhoven: Eindhoven University of Technology.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley: CreateSpace.

VanderPlas, J. (2016). *Python Data Science Handbook*. Sebastopol: O'Reilly Media.

Velaga, N. R., Quddus, M. A., and Bristow, A. L. (2009). "Developing an enhanced weight-based topological map-matching algorithm for intelligent transport systems". *Transportation Research Part C: Emerging Technologies* 17(6), pp. 672–683.

Veltkamp, R. C. (2001). "Shape matching: similarity measures and algorithms". In: *Proceedings International Conference on Shape Modeling and Applications*. IEEE, pp. 188–197.

Ververidis, C. and Polyzos, G. C. (2006). "Location-Based Services in the Mobile Communications Industry". In: *Encyclopedia of E-Commerce, E-Government, and Mobile Commerce*. Ed. by M. Khosrow-Pour. London: Idea Group Reference, pp. 716–721.

Vespignani, A. (2018). "Twenty years of network science". *Nature* 558(1), pp. 528–529.

Vovsha, P. and Bekhor, S. (1998). "Link-nested logit model of route choice: overcoming route overlapping problem". *Transportation Research Record* 1645(1), pp. 133–142.

Wang, W., Jin, J., Ran, B., and Guo, X. (2011). "Large-scale freeway network traffic monitoring: A map-matching algorithm based on low-logging frequency GPS probe data". *Journal of Intelligent Transportation Systems* 12(2), pp. 63–74.

Wardrop, J. G. (1952). "Road Paper. Some theoretical Aspects of Road Traffic Research". *Proceedings of the Institution of Civil Engineers* 1(3), pp. 325–362.

Warshall, S. (1962). "A theorem on boolean matrices". *Journal of the ACM (JACM)* 9(1), pp. 11–12.

Wei, F., Ma, S., and Jia, N. (2014). "A Day-to-Day Route Choice Model Based on Reinforcement Learning". *Mathematical Problems in Engineering* 3(1), Article no. 646548.

Wen, B. (2019). *Wardrop's Principles*. https://www.bowenwen.com/mobilities/wardrop/. last visited on October 7, 2019.

Wen, C.-H. and Koppelman, F. S. (2001). "The generalized nested logit model". *Transportation Research Part B: Methodological* 35(7), pp. 627–641.

Wenk, C., Salas, R., and Pfoser, D. (2006). "Addressing the Need for Map-Matching Speed: Localizing Global Curve-Matching Algorithms". In: *18th International Conference on Scientific and Statistical Database Management (SSDBM'06)*. IEEE, pp. 379–388.

West, D. B. (1996). *Introduction to graph theory*. Upper Saddle River: Prentice Hall.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wickham, H., François, R., Henry, L., and Müller, K. (2019). *dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr. last visited on April 4, 2020.

Wilcoxon, F. (1945). "Individual Comparisons by Ranking Methods". *Biometrics Bulletin* 1(6), pp. 80–83.

Willumsen, L. G. (2000). "Travel Networks". In: *Handbook of Transport Modelling*. Ed. by D. A. Hensher and K. J. Button. Amsterdam: Pergamon. Chap. 10, pp. 165–180.

Wolfermann, A., Friedrich, B., and Fellendorf, M. (2019). "Germany and Austria". In: *Global Practices on Road Traffic Signal Control*. Ed. by K. Tang, M. Boltze, H. Nakamura, and Z. Tian. Amsterdam: Elsevier. Chap. 4, pp. 37–67.

Wong, C.-H. and Tam, Y.-C. (2005). "Negative Cycle Detection Problem". In: *Algorithms – ESA 2005*. Ed. by G. S. Brodal and S. Leonardi. Lecture Notes in Computer Science, vol 3669. Springer, pp. 652–657.

Xu, J., Luo, X., and Shao, Y.-M. (2018). "Vehicle trajectory at curved sections of two-lane mountain roads: a field study under natural driving conditions". *European Transport Research Review* 10(1). Article no. 12.

Yai, T., Iwakura, S., and Morichi, S. (1997). "Multinomial probit with structured covariance for route choice behaviour". *Transportation Research Part B: Methodological* 31(3), pp. 195–207.

Yang, C. and Gidófalvi, G. (2018). "Fast map matching, an algorithm integrating hidden Markov model with precomputation". *International Journal of Geographical Information Science* 32(3), pp. 547–570.

Yao, E., Pan, L., Yang, Y., and Zhang, Y. (2013). "Taxi Driver's Route Choice Behaviour Analysis Based on Floating Car Data". *Applied Mechanics and Materials* 361–363, pp. 2036–2039.

Yuan, N. J., Zheng, Y., Zhang, L., and Xie, X. (2013). "T-finder: A recommender system for finding passengers and vacant taxis". *IEEE Transactions on Knowledge and Data Engineering* 25(10), pp. 2390–2403.

Yuan, N. J., Zheng, Y., Zhang, L., Xie, X., and Sun, G. (2011). "Where to find my next passenger". In: *Proceedings of the 13th international conference on Ubiquitous computing (UbiComp '11)*. ACM, pp. 109–118.

Zhang, D. (2017). "Fastest-Path Computation". In: *Encyclopedia of GIS*. Ed. by S. Shekar and H. Xiong. Cham: Springer, pp. 585–591.

Zhang, H. and Malczewski, J. (2017). "Accuracy Evaluation of the Canadian OpenStreetMap Road Networks". *International Journal of Geospatial and Environmental Research* 5(2). Article no. 1.

Zhang, S., Yao, Y., Hu, J., Zhao, Y., Li, S., and Hu, J. (2019). "Deep Autoencoder Neural Networks for Short-Term Traffic Congestion Prediction of Transportation Networks". *Sensors* 19(10). Article no. 2229.

Zhang, X. (2011). "Support Vector Machines". In: *Encyclopedia of Machine Learning*. Ed. by C. Sammut and G. I. Webb. Boston: Springer, pp. 941–946.

Zheng, Y. (2015). "Trajectory data mining: an overview". *ACM Transactions on Intelligent Systems and Technology (TIST)* 6(3). Article no. 29.

Zheng, Z., Zhang, T., Li, Q., Wu, Z., Zou, H., and Gao, C. (2016). "Curvedness feature constrained map matching for low-frequency probe vehicle data". *International Journal of Geographical Information Science* 30(4), pp. 660–690.

Zhihua, L. and Wu, C. (2005). "A new approach to map-matching and parameter correcting for vehicle navigation system in the area of shadow of GPS signal". In: *Proceedings. 2005 IEEE Intelligent Transportation Systems*. IEEE, pp. 449–454.

Zou, M., Chen, X. M., Yu, H., Tong, Y., Huang, Z., Li, M., and Zou, H. (2013). "Dynamic Transportation Planning and Operations: Concept, Framework and Applications in China". *Procedia - Social and Behavioural Sciences* 96(6), pp. 2332–2343.

Appendix

Attributes in the routes dataset

Each route in the routes dataset is provided with the attributes listed and explained below. The keyword ”trip” specifies that the attribute is shared between the actual and the alternative routes for the specific trip. The keyword ”route” indicates that the attribute is route-specific.

`city_long`, *varchar(13)*
Full name of the city in which the route is located.

`city_short`, *char(3)*
Abbreviated name of the city in which the route is located.

`trip_id`, *char(10)*
Trip identifier containing the abbreviated city name and numeric digits.

`trip_starttime`, *int*
Trip's start timedate in unix-time.

`trip_od_distance`, *numeric(7,2)*
Length of the shortest route between origin and destination in metres. For shortest routes, this is equal to the route's length.

`route_type`, *varchar(13)*
Descriptor whether it is an actual, shortest, fastest, or lowest intersections route.

`route_length`, *numeric(7,2)*
Total length of the route in metres.

`route_shared_length`, *numeric(7,2)*
Length of the route that was shared with the actual route in metres. For actual routes, this is equal to the total route length.

`trip_actual_duration`, *smallint*
Total actual duration of the trip in seconds derived from the timestamps in the raw data.

`route_optimal_duration`, *smallint*
Theoretical duration of the route in seconds. The duration is calculated based on maximum speed limits which means, that it displays the time a vehicle would need to cover the given route if it could drive with the maximum speed allowed and without slowing down on intersections, traffic lights, or crossings.

`route_intersections`, *smallint*
Total number of intersections on the route.

`route_traffic_signals`, *smallint*
Total number of traffic signals on the route.

`route_turns`, *smallint*
Total number of turns on the route.

`route_betweenness_centrality`, *numeric(11,10)*
Average EBC of all network edges the route travels.

`route_psl`, *numeric(5,2)*
PSL between the route and its corresponding actual route. Actual routes always have a PSL of 100 % (see Equation 2.2).

`route_pld`, *numeric(5,2)*
PLD between the route and its corresponding actual route. Positive values mean that the actual route is longer and negative values indicate that the actual route is shorter than the alternative route. Actual routes always have a PLD of 0 % (see Equation 2.3).

`route_ptd`, *numeric(5,2)*
PTD between the route and its corresponding actual route. Positive values mean that the actual route takes longer and negative values indicate that the actual route is faster than the alternative route. Actual routes always have a PTD of 0 % (see Equation 2.4).

`route_pid`, *numeric(5,2)*
PID between the route and its corresponding actual route. Positive values mean that the actual route includes more intersections and negative values indicate that the actual route visits less intersections. Actual routes always have a PID of 0 % (see Equation 2.5).

`route_frechet`, *numeric(7,2)*
FD distance between the route and its corresponding actual route in metres. Actual routes always have an FD of 0 metres.

`route_turns_r`, *smallint*
Number of left turns on the route.

`route_turns_l`, *smallint*
Number of right turns on the route.

`route_turns_sharp_r`, *smallint*
Number of sharp right turns on the route.

`route_turns_sharp_l`, *smallint*
Number of sharp left turns on the route.

`route_intersections_3`, *smallint*
Number of intersections with 3 intersecting streets.

`route_intersections_4`, *smallint*
Number of intersections with 4 intersecting streets.

`route_intersections_5`, *smallint*
Number of intersections with 5 intersecting streets.

`route_intersections_6`, *smallint*
Number of intersections with 6 intersecting streets.

`route_intersections_larger6`, *smallint*
Number of intersections with more than 6 intersecting streets.

`route_living_street`, *numeric(5,2)*
Percentage of the route travelling on streets where pedestrians have legal priority over traffic and children are allowed to play on the street.[48]

`route_residential`, *numeric(5,2)*
Percentage of the route travelling streets serving as access to housing.[48]

`route_unclassified`, *numeric(5,2)*
Percentage of the route travelling streets at the lowest level of the street network. Note that this is not a placeholder for streets without a classification.[48]

`route_tertiary`, *numeric(5,2)*
Percentage of the route travelling on minor streets.[48]

`route_secondary`, *numeric(5,2)*
Percentage of the route travelling on medium streets.[48]

[48] For further information, visit https://wiki.openstreetmap.org/wiki/Key:highway.

`route_primary`, *numeric(5,2)*
Percentage of the route travelling on major streets.[48]

`route_trunk`, *numeric(5,2)*
Percentage of the route travelling on high performance streets.[48]

`route_motorway`, *numeric(5,2)*
Percentage of the route travelling on restricted access major divided highways.[48]

`route_other`, *numeric(5,2)*
Percentage of the route travelling on streets whose type is not defined by the attributes above.[48]

`trip_origin_lon`, *numeric(8,3)*
Origin longitude in WGS84 coordinates.

`trip_origin_lat`, *numeric(8,3)*
Origin latitude in WGS84 coordinates.

`trip_destination_lon`, *numeric(8,3)*
Destination longitude in WGS84 coordinates.

`trip_destination_lat`, *numeric(8,3)*
Destination latitude in WGS84 coordinates.

`trip_origin`, *geometry*
Point geometry of the trip origin in hex-encoded EWKB format in WGS84 coordinates.

`trip_destination`, *geometry*
Point geometry of the trip destination in hex-encoded EWKB format in WGS84 coordinates.

`route_geom`, *geometry*
LineString geometry of the route in hex-encoded EWKB format in WGS84 coordinates.

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.

Winterthur, June 28, 2020

Tim Waldburger