Technische Universität München
Department of Civil, Geo and Environmental Engineering

Do Taxi Drivers Choose the Shortest Routes?

A Large-Scale Analysis of Route Choice Behavior of Taxi Drivers using Floating Car Data in Shanghai

Junyan Li

A thesis presented for the degree of
M.Sc. Cartography
2017

Supervisors
Juliane Cron M.Sc. (Technische Universität München)
Univ.Prof. Mag.rer.nat. Dr.rer.nat Georg Gartner (Technische Universität Wien)
Dr. Haosheng Huang (University of Zurich)

Submission: September, 2017

DECLARATION OF AUTHORSHIP

I hereby declare that the submitted master thesis entitled *Do Taxi Drivers Choose the Shortest Routes? A Large-Scale Analysis of Route Choice Behavior of Taxi Drivers using Floating Car Data in Shanghai* is my own work and that, to the best of my knowledge, it contains no material previously published, or substantially overlapping with material submitted for the award of any other degree at any institution, except where acknowledgment is made in the text.

Munich, 21st of September 2017

Junyan Li

Abstract

Understanding taxi drivers' route choice behavior is still a theoretical and empirical challenge. For this end, tradition approaches construct rational choice models and predict deterministically the outcome from a finite set of choices. Although useful, this modeling approach omits a large proportion of elements such as number of left turns, road type, travel distance and travel time that can influence taxi drivers' choices in real world. Under this context, the case of Shanghai is of particular interest because it is composed of an enormous amount of taxi drivers' Floating Car Data and includes a large area of traveling activities. In contrast with traditional formal modeling approach, this thesis empirically studies the influence of selected features on the route choice behavior of taxi drivers with a case study in Shanghai. This large-scale analysis uses more than 650 million GPS points and around 3.6 million valid routes from the Floating Car Data. The thesis resolves the questions about whether the shortest routes determine the choice of taxi drivers, and to what extent other potential elements may impact decision making of taxi drivers. Utilizing Floating Car Data in Shanghai, the analysis proceeds first with data pre-processing step which reordered, cleaned and extracted data of interest, as well as using map matching to integrate GPS points to the road network for further analysis. Then an overall pattern of taxi service activities is presented which includes travel areas, frequencies, durations, etc. Thereafter, two types of analyses are conducted. The first type is comparison analysis which compares the real and the shortest routes in terms of both route-based features and segment-based features. The second type is preference analysis that formulates scenarios and compares different features in each scenario once at a time. While real routes to some extent overlap with the shortest routes with an average value of 46%, the results show that other features, especially the road type, road usage frequency, speed limit and observed travel speed, have impacts on the route choices. In the general situation, the best scenario in combination with travel distance and road usage frequency predicts 53.56% of real routes compare to 46.61% in the shortest routes. In long distant trips, the scenario with road type and travel distance predicts 51.79% of real routes compared to 13.25% in the shortest routes. What comes as surprise is the insignificance of number of turns in the results in contrast to previous scholarly findings.

Keywords: taxi driver, route choice behavior, Floating Car Data, Shanghai

Contents

List of Figures

List of Tables

Acronyms

DBSCAN Density-Based Spatial Clustering of Applications with Noise. 55

FCD Floating Car Data. 2

GPS Global Positioning System. 3

HMM Hidden Markov Model. 19

OD Origin and Destination. 28

PLD percentage of length difference. 31

PSL percentage of shared length. 31

RDI Route Directness Index. 38

1 Introduction

1.1 Context of study

Route choice is one of the fundamental features of human wayfinding (Golledge, 1995), and understanding route choice behavior of drivers is critical to administrative purposes such as urban planning, traffic management, and infrastructure construction — and it is also of keen interest to the research and investigation on epidemiology and social security. With the expansion of road network in recent years, it increases the number of route alternatives from one place to another, and thereby increases the complexity of route behavior analysis.

Following the shortest route is traditionally believed to be one of the major features of human route choice behavior, because everything else equal, the shortest route has the minimum travel time and induces the least cost with the lowest possibility of congestion and traffic accidence (Golledge, 1995). Compared to ordinary drivers, taxi drivers are more experienced in wayfinding, and are more familiar with road conditions, possible alternative routes, and traffic situation, and therefore are more likely to choose the route with the shortest distance (Yao et al., 2013). However, Golledge (1995) also points out that besides the shortest routes, other criteria such as fewest turns may also have an impact. However, it is not well established yet to what extent taxi drivers follow the shortest route in presence of multiple factors.

1.2 Motivation and problem statement

In spite of the seemly simplicity when it is first addressed, the pooling and interaction of individual taxi drivers within a huge traffic network adds to the complexity of the problem. First of all, taxi drivers are humans whose decision making process consists of both rational and irrational components (de Carvalho et al., 2016). Therefore, it makes accurate prediction of route selection difficult. Second, while theories predict that multiple equilibriums that result in choosing different routes (Smith, 1979, 1983, 1984), it is hard to justify the theory prediction without large-scale observable data. Finally, observable data of taxi drivers' route choice is always GPS-based point data. The data needs to be pre-processed and integrated with existing road network in order to make valuable inference out of it (Guo and Huang, 2016).

Due to the availability of Floating Car Data (FCD), current technology enables scholars to utilize large scale data with accurate temporal and positional information to tackle the problem of taxi driver route choice behavior. Utilizing FCD, this thesis aims to show to what extent real routes mimic the path of the shortest routes and whether selected factors influence taxi drivers to deviate from the shortest routes.

1.3 Research questions, approach and significance

The research questions in this thesis are as follows.

- To what extent do real routes overlap with the shortest routes, and how much longer are real routes compared with the shortest routes?
- Are other factors such as speed limitation, road type, number of left turns, number of intersections, road usage frequency impacting the deviation of real routes from the shortest routes?

To systematically study the questions at hand, this thesis adopts an empirical approach by using massive FCD to examine the route choice behavior of taxi drivers. Shanghai is of particular interest because it has a huge traffic flow of taxi drivers and the activities of drivers are among the busiest around the world. There are at least two reasons to use FCD to examine route choice behavior of taxi drivers. First, compared to survey data, FCD reveals much more information such as travel speed, real route choice and travel direction. Second, FCD stores information of massive number of taxi drivers with relatively cheap costs. Thus, the data allows to systematically examine the route choice behavior of taxi drivers.

The importance of this study is threefold: First, most of the existing studies focus on the behavior of ordinary drivers, and there is relatively little information about the driving behavior and choice pattern of taxi drivers. This thesis contributes to the literature by adding additional examination of taxi driver route choice with large scale data; Second, it will provide cross-validation for previous studies by adding another sample from Shanghai; Finally, this study will help, in addition to scholars, both individual taxi drivers and taxi companies to better understand the patterns of route choice.

1.4 Structure of the thesis

A literature review in the following chapter is given to examine the theoretical and technical backgrounds of the existing explanations of route choice behavior. In the third chapter, the data and study area will be presented including the area and period of study, the number of recorded samples and the structure of the data. Real routes will be extracted from observed Global Positioning System (GPS) points through pre-processing, which includes origin-destination identification, filtering of irrelevant and erroneous points and map matching in the chapter four. After pre-processing, the chapter five will provide a preliminary overview of the patterns comparing between real routes and shortest routes. Afterwards, a family of comparison analysis will be conducted to investigate potential factors that may influence taxi drivers' route choice behavior. Based on the comparison analysis, k-shortest-path-based preference analysis will be used to examine to what extent do these factors impact the final choice and whether their impacts differ in certain type of trips. The last section concludes and provides an outlook for future research. Figure 1 illustrates the main work flow in data processing and analysis.

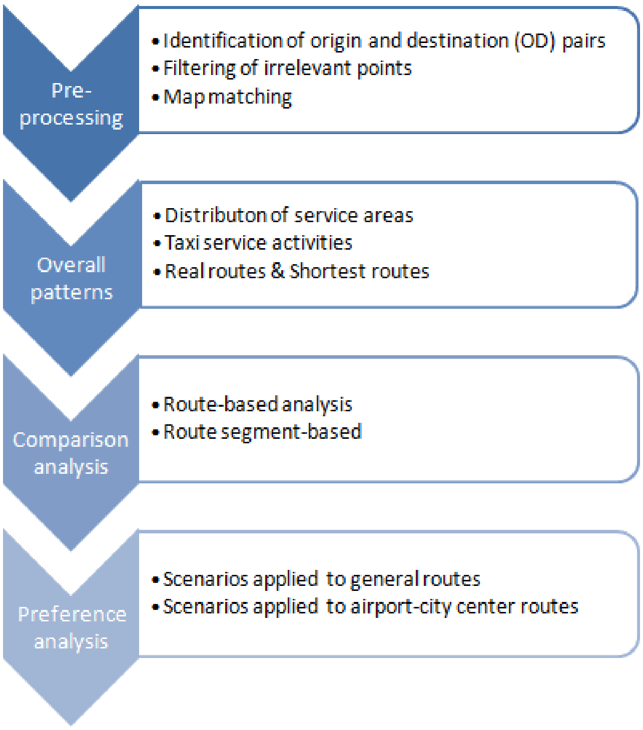

Pre-processing
• Identification of origin and destination (OD) pairs
• Filtering of irrelevant points
• Map matching
Overall patterns
• Distributon of service areas
• Taxi service activities
• Real routes & Shortest routes
Comparison analysis
• Route-based analysis
• Route segment-based
Preference analysis
• Scenarios applied to general routes
• Scenarios applied to airport-city center routes

Figure 1. Work flow

2 Theoretical and technical background

Investigation of route choice behavior has its long tradition. Through decades of research, scientists have proposed both theories and methods to better understand the factors influencing people's route choice. The next sections introduce most relevant theories and recently developed methods that paved the way for the following analysis.

2.1 Route choice behavior

Existing literature has produced a rich source of information on the mechanism of route choice and how multiple factors generate and constrain the possible choice set. As a sub-branch of route choice study, research on drivers' route choice behavior has also been widely investigated from both theoretical and methodological perspectives. Theoretically, a family of formal models from a game-theoretical perspective to analyze individual risky choice and its collective implication have been developed (Arnott et al., 1992; Wardrop, 1952; Smith, 1979, 1983, 1984). Methodologically, multiple approaches including simulation (Agrawal et al., 2016), survey (Razo and Gao, 2013; Yang et al., 2015), and observed data analysis (Gan et al., 2010; Papinski and Scott, 2011, 2013; Sun et al., 2014; Tang et al., 2016) have been adopted in revealing drivers' route choice patterns.

The following sections describe different factors influencing taxi drivers' route choice.

2.1.1 Rationality and bounded rationality

Rationality is the fundamental assumption in many models of drivers' route choice behavior (Miller-Hooks and Mahmassani, 2000; Razo and Gao, 2013; Smith, 1979, 1983, 1984). It assumes that a driver calculates costs and benefits given a set of alternative routes, and will choose the route that gives him/her the highest marginal utility. The idea can be traced back to Von Neumann and Morgenstern (1944) who show that in face of several risky choices, a person will choose the one that maximizes his/her expected utility. The pioneering work by Wardrop (1952) discusses some theoretical framework to include the aspects of rational choices into analyzing traffic behavior. By developing further the concept to the analysis of transportation network, Smith (1979, 1983, 1984) investigates the existence, uniqueness and stability of traffic equilibriums based on collective route choice behavior. More recent studies adopt the concept and apply it to drivers' route choice behavior (He and Peeta, 2016; Jiang et al., 2014; Yang and Zhang, 2009). However, in these studies, the components of drivers' utility function is predefined by the researchers, and it is unclear to what extent these components affect drivers' choices.

Irrationality component of drivers' evaluation also adds to the complexity of the problem. Following Smith's works, refined versions of equilibrium analysis have been proposed by recent scientists to account for bounded rationality (de Carvalho et al., 2016; Guo and Huang, 2016). It is assumed that drivers are indifferent among a set of alternatives when their differences are small. Therefore, the minimal-cost route in theory may not be the actual route been chosen. This idea of bounded rationality adds insights into analysis of taxi drivers' route choice behavior and indicates the variation of drivers' real choices.

2.1.2 Travel distance

Indeed, there are multiple structural factors influencing drivers' choices and the extent of their influences varies from vehicles to vehicles. For taxi drivers, the most prominent factor is probably the travel distance, since unlike other factors, it exerts the most direct influence on the travel budget. In particular, the travel distance is such an essential factor that it is the main component included in the early static analysis (Wardrop, 1952). By investigating customer-searching behavior of taxi drivers in Beijing, Tang et al. (2016) show in their empirical analysis of GPS-based data that taxi drivers prefer routes with shorter distances and less travel time. While it is largely unknown whether on-service taxi drivers have the same preferences as those on searching, researchers generally include distance and time into their analysis (Papinski and Scott, 2011, 2013; Ramaekers et al., 2013; Yang et al., 2015). Using FCD from Shenzhen, Sun et al. (2014) show that drivers' route choice is influenced by multiple factors especially the travel distance. However, recent research also shades doubt on the validity of shortest-route prediction. In particular, GPS-based data show that there is high deviation of real routes from the corresponding shortest routes (Ciscal-Terry et al., 2016; Manley et al., 2015).

2.1.3 Travel time and congestion

Drivers are sensitive to travel time and also respond to congestions (Prato et al., 2014; Prato and Rasmussen, 2016). Because congestion changes the accessibility of the points along the congested route (Li et al., 2011), taxi drivers will try to avoid the routes with high probability of congestion (Arnott et al., 1992; Palma and Rochat, 1999; Tadic et al., 2007). And also because the traffic situation is time-dependent and varies across hours within a day (Palma and Rochat, 1999), taxi drivers have incentives to act differently in response to differences in the likelihood of congestion and low speed traveling.

2.1.4 Road type

Road type is another important factor in influencing taxi drivers' route choice (Manley et al., 2015). Different road types have varying levels of quality and traffic flows. Taxi drivers may prefer the road with high speed limit and low possibility of congestion. With a case study on minicab routing in London, Manley et al. (2015) find that route features such as direction, deviation and potentials for congestion significantly determine the drivers' final choices.

Existing theoretical framework lends much insights into studying taxi drivers' route choice behavior. In particular, formal analysis provides a theoretical framework under which taxi drivers' rational and irrational components together result in the real route choices. However, current research lacks solid empirical evidence to confirm the previous arguments especially on how and to what extent the structural factors, including travel distance, timing, congestion and road types, influence the real behavior of taxi drivers in route choice. Thus, to investigate in detail of taxi drivers' choices, it is necessary to gather and handle empirical evidence that may either give support or go against the previous theoretical argument.

2.2 Floating Car Data

Even though simulation- and survey-based studies have output delightful results to understand route choice behavior, it is still questionable whether they are able to mimic the real world situation with great complexity. In addition, survey-based techniques are often very expensive and time-consuming, and hard to apply in large-scale studies. In contrast to survey data, GPS-based FCD is one of the reliable traffic data sources (Paulin and Bessler, 2013) and it provides detailed trip information such as origin and destination (OD), operation time, speed and direction and offers the possibility to accurately observe the actual routes drivers have chosen. Due to budget and coverage advantages, FCD allows for large-scale analysis which helps to reduce the survey-based selection bias.

Increasing speed of vehicles and continuously expanding road networks requires technologies that are able to gather and explore large scale information in a real-time manner. GPS based FCD is regarded as a such a reliable and cost-effective way to gather accurate traffic information from large scale road networks (De Fabritiis et al., 2008). Unlike fixed-point sensors such as radars and cameras, the GPS devices installed on floating cars are able to provide near-real time information including position, speed, direction, and time, among other kinds of information.

In order to collect data from moving vehicles, the GPS receiver installed on-board must be locked on to the signals from at least three satellites and a wireless communication device will send and receive signals that will be transferred into interpretable information (Hofmann-Wellenhof et al., 2012).

Since the availability of FCD, it has been used in a wide range of applications. At the macro level, FCD is used to construct real time traffic information system as well as inference of mobilities and congestions. Schafer et al. (2002) use FCD to reconstruct the traffic information systems in Berlin, Nuremberg and Vienna, so that the reconstructed system represents near-complete coverage of real road network and is able to reflect the real time traffic conditions. Similar works also include Kerner et al. (2005), Huber et al. (1999) and Gühnemann et al. (2004). FCD is also used to infer accessibility and mobility of road networks so that transportation planners are able to map the accessibility of certain points of interest from the constrained network (Li et al., 2011).

At the micro level, FCD is used as a way to predict various information from individual drivers. In the area of logistic planning, FCD is used to predict travel time for both short- and long-range trips of logistic vehicles (Simroth and Zähle, 2011). It can also be used to explore individual driver behaviors. Ciscal-Terry et al. (2016) use the low-frequency GPS-based FCD from Italy to study the drivers' route choice behavior in response to territory of the traveling area. Sun et al. (2014)

use FCD of Shenzhen to analyze commuting driver behavior in urban areas. Using FCD from Greater Copenhagen area, Prato et al. (2014) and Prato and Rasmussen (2016) show that drivers are more responsive to travel time and congestion during rush hours in a day and therefore are more likely to avoid the road with high congestion probabilities during that time. Tang et al. (2016) use FCD of vacant taxis in Beijing to investigate taxi customer searching behavior. And similarly, based on FCD of taxi drivers in Beijing, Yao et al. (2013) find that taxi drivers are more likely to choose the route with faster speed, less left turns and more express ways. With a case study of London minicab, Manley et al. (2015) compare shortest path and anchor-based route choice and show that anchors (route features) impact minicab routing.

2.3 Map matching

Map matching is the process of integrating the GPS-based FCD with spatial road network data to identify the location of a vehicle on the road network. For this purpose, a variety of algorithms were developed to improve integrating accuracy and processing efficiency. Quddus et al. (2007) provide an overview of state-of-the art algorithms to deal with map matching:

The first and most frequently used algorithms, which include point-to-point matching (Bentley and Maurer, 1980; Bernstein and Kornhauser, 1998), point-to-curve matching (White et al., 2000) and curve-to-curve matching (Phuyal, 2002; White et al., 2000), are based on the geometric distances between the GPS points and the road network. However, these geometric-based algorithms are sensitive to road network density and are unstable with presence of outliers (Quddus et al., 2007).

The second type of algorithms take into account of the topological features of road networks such as connectivity, contiguity and curvature (Quddus et al., 2007). In combination with geometric features, topological analysis commonly adopt a weighted scheme that utilizing available information including road turn, travel speed, travel direction, and similarity between the vehicle trajectory and the road network (Greenfeld, 2002; Quddus et al., 2003; Yu, 2006). And therefore, the weighted topological map-matching algorithms always have better performance than those solely based on geometric features (Quddus et al., 2007).

The third type of algorithms derive confidence intervals surround the GPS points and match accordingly the road segments with the highest confidence level (Ochieng et al., 2003; Zhao, 2010). However, this type of algorithm cumulates errors with every construction of confidence interval and additional criteria are needed to complement the accuracy loss (Quddus et al., 2007).

More advanced algorithms also incorporate researches on belief theory, fuzzy logic and Bayesian analysis (El Najjar and Bonnifait, 2005; Krakiwsky et al., 1988; Kim, 1998; Newson and Krumm, 2009; Pyo et al., 2001), and are reported to perform better than aforementioned algorithms (Quddus et al., 2007). The most recent development of map-matching algorithms regards the GPS points as a serial revealing of latent positions on the road network that transfer between different states from previous to current time (Newson and Krumm, 2009). Therefore, the probability of a set of GPS points located on certain combinations of road segments can be calculated from a Hidden Markov Model where the current position only depends on the nearest early position and a transferring probability matrix. This algorithm regards the GPS records as a serial connected set instead of uncorrelated individual points, and takes the connectivity between different points into account. Consequently, the algorithm is able to derive locations on road network in a consistent way that won't be impacted too much by outliers.

2.4 Algorithms for finding the shortest route

Finding the shortest route poses another challenge to studying route behaviors in large road network. In recent years, there have been rapid development of algorithms for finding and comparing shortest routes with multiple criteria that can be further added. The following explores existing algorithms from the simplest shortest path algorithm to k-shortest path algorithm and to other refinements.

2.4.1 Shortest path

The problem to find the shortest path by constructing a tree with minimum total length between n node was first proposed by Dijkstra (1959). And one of the initial algorithms that was clearly stated for finding the shortest route was introduced by Floyd (1962) who used recursive loops to identify the shortest path from node i

to j. In detail, the algorithm initializes an storing array $m[i, j]$ corresponding to the length of a direct link between i and j. Then it loops through 1 to N—the number of nodes in the network—to find the path that has the lowest cumulative distances.

The algorithm provides the basis for finding the shortest route based on geometric distances. However, the shortage of the algorithm is also apparent. It is unable to handle calculation of shortest path that are based on other criteria except for distances. Another limitation of this algorithm is that only a single route will be produced instead of a set of them, which are commonly required for comparison analysis.

A partial improvement was done by Ziliaskopoulos and Mahmassani (1993) who introduced time-dependence into calculation of shortest routes so that the algorithm is capable of handling real-time data from traffic records. Jonker and Volgenant (1987) also demonstrate a new algorithm that implements new initialization routines for linear assignment problems.

2.4.2 K-shortest path

Given the need to compare a set of routes between two end nodes, the k-shortest path algorithm was developed to find the specified number of shortest routes from the road network in an order from 1 to k. Hoffman and Pavley (1959) propose a looping solution to find the k shortest path, but the algorithm is computationally intensive when the value of k is big and the scale of the network is large. Yen (1971) presents an efficient algorithm that is able to calculate k loop-less paths that have the shortest total length in a network. The algorithm is fast and its computational time only increases linearly with the increasing value of k. Eppstein (1998) achieves further computational improvement over the algorithm and Wilson (1996) introduces the algorithm with additional weighting scheme.

The above theoretical and methodological introductions based on existing literature provide an overview of factors influencing individual route choice behavior, as well as the state-of-the-art methods for investigating the route choice problem. The following chapter begins with an overview of the data and the study area after which a series of analyses are conducted.

2.5 State of the art and research gaps

While the above sections describe the relevant theories and technologies, this section discusses in detail the most relevant research on taxi drivers' route choice behavior. Drivers' route choice have been widely studied from different perspectives. In particular, advanced technologies such as simulation (Agrawal et al., 2016; Gelenbe and Sakellari, 2012; Shang et al., 2016; Yan et al., 2006), stated preferences (Razo and Gao, 2013; Yang et al., 2015), complicated game theoretical models (Smith, 1979, 1983, 1984; Wardrop, 1952) are used. However, till now there is only a fraction of literature on route choice behavior focusing on taxi drivers. It includes the works by Yao et al. (2013), Sun et al. (2014) and Manley et al. (2015).

Yao et al. (2013) investigate taxi drivers' route choice with FCD in Beijing. They study the influence of road network conditions and traffic status on the choice of taxi drivers, and they find that taxi drivers tend to choose the route with faster travel speed, less left turns and more proportion of highways. However, they only use a very small subsample of taxi trips (in total 221 trips) that are randomly chosen from the FCD. Such small number of trips limits the application of their findings to the general situations with more trips and much larger networks. In addition, because only a small number of factors are analyzed, their study cannot rule out the influence of factors such as observed travel time.

Sun et al. (2014) utilize FCD in Shenzhen to study the taxi drivers' route choice behavior. They distinguish between the shortest route and the fastest routes, and specify eight scenarios to quantify the influence of various potential factors. Their results show that travel distance, travel time and road preference have high influence on taxi drivers' route choices. Even though they have much larger samples than Yao et al. (2013), the total number of trips is around 4000, which is much smaller than the number of trips in real-world operations.

Manley et al. (2015) study the minicab trips in London also with FCD. They increase the number of trips under study to around 700,000 with about 3000 drivers, and establish what they call anchor-based analysis that combine different road features together in predicting the route choice. Compared to shortest path prediction, their results show that anchors like travel time, turns and road category have higher impacts on minicab drivers' route choices. So far, their study is the most sophisticated one on the taxi drivers' route choice behavior. However, it also has limitations: First, they are able to general only a single shortest route at a time to minimize the cost function, which is less desirable than general multiple routes that connect between the origin and the destination. Second, instead of analyzing the whole London network, the authors only select cases in smaller regions, which may not be representing the whole city. Third, the number of trips in their study is still small than real situations.

Borrowing insights from the above studies and considering the limitation of them as well, this thesis addresses the problem of taxi drivers' route choice with a large-scale dataset that contains more than 3,600,000 valid taxi trips. A number of influencing factors are identified and specified into scenarios to examine their relative influence on the taxi drivers' route choice. The details of the analysis will be described in the following chapters.

3 Data and study area

To give an overview, this chapter introduces the area and time of study, the number of records as well as the data structure.

3.1 Area of study

Shanghai is of particular interest because as the economic center of China, it has the busies traffic flows around the world. Among other traffic means, taxis are frequently used by local residents and visitors for commuting between different areas of the city. Therefore, it provides us a rich set of information of taxi drivers' routine around the clock. In particular, the GPS-based FCD records the points along the trajectories of taxi drivers collected by a fleet of vehicles. The data provides valuable information for inferring the movement of taxi drivers in an accurate and cheap manner. Figure 2 shows the area of study with the road network, major districts and two airports.

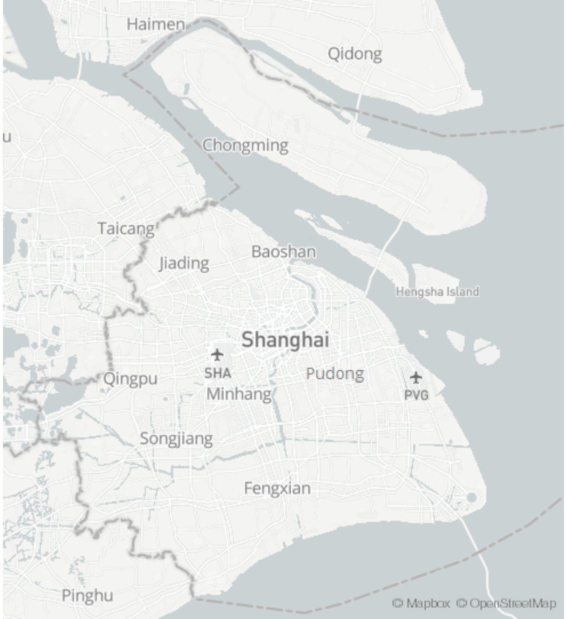


Figure 2. Topological map of Shanghai with the road network districts, and airports

3.2 Time period of study and data

To present representative results, the period of study covers two weeks from June 15th (Tuesday) to June 28th (Monday), 2010. During this period, all records of taxi FCD in Shanghai with passengers on board are used for analysis [1]. The total number of taxi drivers in the sample is 7168.

The data used in this study contains the information on the longitude and latitude of points along the trajectory as well as other information such as time, directions, and instant velocities. Table 1 summarizes the main attributes of the data from the single day, and shows the examples extracted from the data. In addition to spatial and temporal information, the taxi id and the passenger indicator help to distinguish different operating taxis and identify the services status of taxi drivers.

Column	Attributes	Example
date	Date of FCD Data	20100615
taxiid	taxi id, consists 5 letters	10003
lon	longitude of current location	121.529558
lat	latitude of current location	31.23496
velocity	instant speed	45.5
direction	driving orientation	155
	valid number from 0 to 359 degree	
passenger	if the taxi is with passenger or not	1
	No - 0, Yes - 1	
time	exact time of current location	2010-06-15 00:00:07

Table 1. Data structure of FCD data in Shanghai

Table 2 shows the summary of the whole dataset including the information about the numbers of FCD points and taxi driver routes over the 14 days. On average, there are around 50 million GPS points and more than 250 thousands valid on-service trips each day. On-service trips mean the routes chosen by taxi drivers when there are passengers on board. In total, there are around 650 million GPS points and around 3.6 million valid routes.

[1] The data was shared by Prof. Chun Liu from Tongji University.

Date	FCD Points	Routes
20100615	55,083,889	275,958
20100616	56,300,003	268,583
20100617	51,978,838	272,594
20100618	47,390,204	278,074
20100619	50,444,522	290,734
20100620	51,259,864	258,392
20100622	47,272,895	271,188
20100623	48,146,691	280,074
20100624	48,267,510	283,901
20100625	48,598,111	295,160
20100626	49,348,793	296,208
20100627	50,001,994	273,538
20100628	44,335,553	273,417
SUM	648,428,867	3,617,821

Table 2. Valid number of points and routes for each day included in the analysis

After having the original data source, the next chapter will pre-process the data to get rid of erroneous records, identify origin and destination pairs and match the GPS points to the road network.

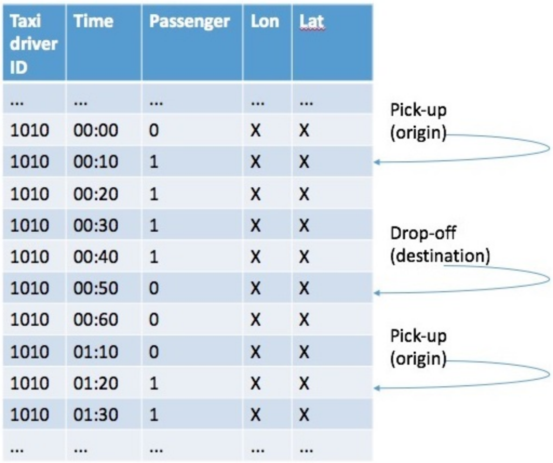
4 Pre-processing

The pre-processing stage consists of identification of origin and destination pairs, filtering of irrelevant points and map matching process. The following sections will describe each of them in detail.

4.1 Identification of origin and destination pairs

FCD only records GPS points by time and the format of the data is not ready for immediate analysis. Therefore, it is necessary to sort out the set of points representing a complete route from the data. To do so, it is required to identify the origin and destination (OD) pairs of on-service trips from original data.

To identify OD pairs from original FCD, a searching algorithm that vertically loops through the data was adopted. Figure 3 shows the mechanism of the algorithm. First of all, data for each day is sorted by taxi driver id and time to establish an looping environment. Second, the passenger variable (1 for passenger on board and 0 for vacancy) is used as the index for identifying the change of status from 1 to 0 or from 0 to 1. If the status change is from 0 to 1, the beginning point with passenger value of 1 is identified as the pick-up (origin) point, and vice versa to identify the drop-off point. Only the records between the pick-up and the drop-off point will be used as relevant data. In other words, the data without passengers on board will be deleted from the original data.



Taxi driver ID	Time	Passenger	Lon	Lat
...	...	...	...	...
1010	00:00	0	X	X
1010	00:10	1	X	X
1010	00:20	1	X	X
1010	00:30	1	X	X
1010	00:40	1	X	X
1010	00:50	0	X	X
1010	00:60	0	X	X
1010	01:10	0	X	X
1010	01:20	1	X	X
1010	01:30	1	X	X
...	...	...	...	...

Pick-up (origin)
Drop-off (destination)
Pick-up (origin)

Figure 3. Searching alogrithm to identify OD pairs

4.2 Filtering of irrelevant points

After conducting the OD searching, the resulting data is still not clean enough for map matching and the following analysis. There are four reasons for it. First, a taxi driver may forget to turn off the passenger-on-board indicator or mistakenly turn on it when the taxi is off service and not moving. Second, taxi drivers may mistakenly turn off or turn on the passenger-on-board indicator in a very short time. This results in the change of recorded service status within seconds even though the real status does not change. Third, the taxi drivers may accidentally or intendedly drive off the roads which results in unmatched routes in the sample. Finally, the normal records of direction for each GPS point should lie between 0 and 360 degrees. However, the GPS data may produce errors when records of directions are larger than 360 degrees.

In order to further filtering the irrelevant points due to the aforementioned reasons, the following cleaning procedure was implemented. First, to eliminate points with wrong status, the trips with average speed less than 1 km/h were deleted. Second, to deal with immediate service status change, routes with length less than 1 km were removed. Third, to cope with off-path drivings, the routes that cannot be map-matched were deleted. Finally, the recordings with direction larger than 360 degrees were deleted.

4.3 Map matching

Map matching forms the process of associating a point to a route segment and constructing consecutive trajectories that are essential for assigning attributes such as length and speed to the operating taxis. Because the recorded GPS data are concrete points with random errors from factors like weather and satellite visibility, for analyzing taxi drivers' route choice behavior, it is necessary to perform map matching before analyzing the route characteristics.

The map matching algorithm adopted in this study is based on Hidden Markov Model (HMM) (Newson and Krumm, 2009). The idea behind the algorithm is that: the most likely sequence of routes are chosen by taking into consideration of both the distance between points and routes, and between the consecutive points. In detail, HMM takes into account of both the recorded GPS locations, the time stamp of recorded points and the connectivity of the road network. It smooths the noisy data by calculating the path with the highest likelihood from time sequence of the recorded GPS data.

In detail, the HMM map matching calculates the emission probability of an observed point location given the real located segments, and the transition probabilities from one stage to another. At time t, the emission probability of an observed point location z_t given the true located segment r_i is given as

$$p(z_t|r_i) = \frac{1}{\sqrt{2\pi}\sigma_z} e^{-\frac{1}{2}\left(\frac{||z_t - x_{t,i}||_{\text{greatcircle}}}{\sigma_z}\right)^2}$$

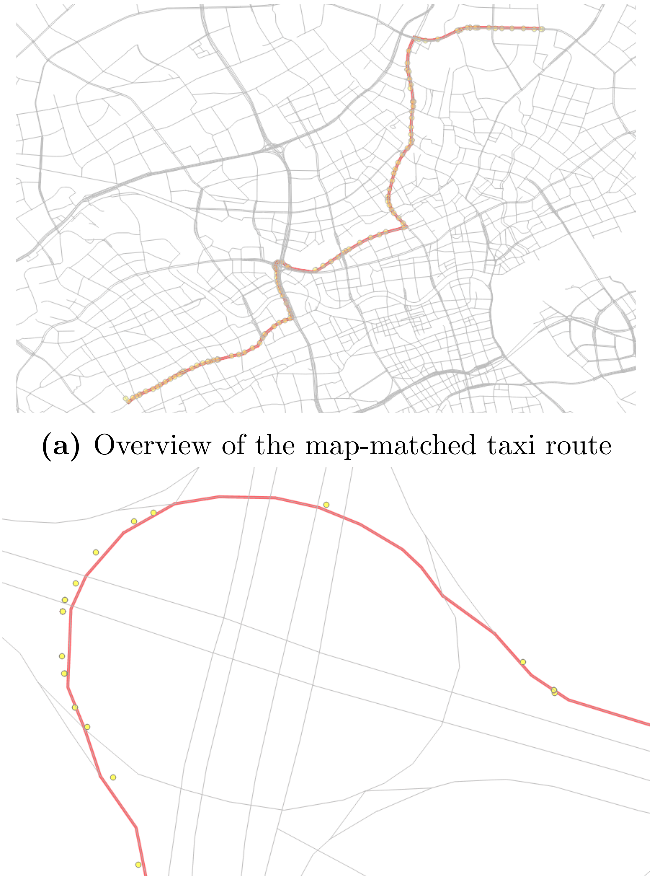
which is a Gaussian distribution with standard deviation σ_z estimated directly from the data.

And the transition probability is modeled as an exponential distribution

$$\begin{aligned} p(d_t) &= \frac{1}{\beta} e^{\frac{-d_t}{\beta}} \\ \text{in which} \quad d_t &= \text{abs}\left(||z_t - z_{t+1}|| - ||x_{t,i} - x_{t+1,j}||\right) \end{aligned}$$

where i and j are the true route segments. The Viterbi algorithm was used to calculate the optimal path by dynamic programming (Newson and Krumm, 2009).

Figure 4 in addition illustrates an example of a taxi route connecting the pickup and the drop-off locations. The panel (a) shows the whole matched route and the panel (b) shows the details of the GPS points and the corresponding routes. The dots represent the points recorded by GPS devices installed on the taxis, and the lines connecting the dots represents the matched routes. From observation, the routes are matched pretty good with the observed GPS points.



(a) Overview of the map-matched taxi route

(b) Details of the map-matched result

Figure 4. Example of a map-matched taxi route in Shanghai

The pre-processing stage delivers the cleaned data that are ready for exploration and analysis. Before going into detailed analysis, the following chapter will first examine the overall patterns .

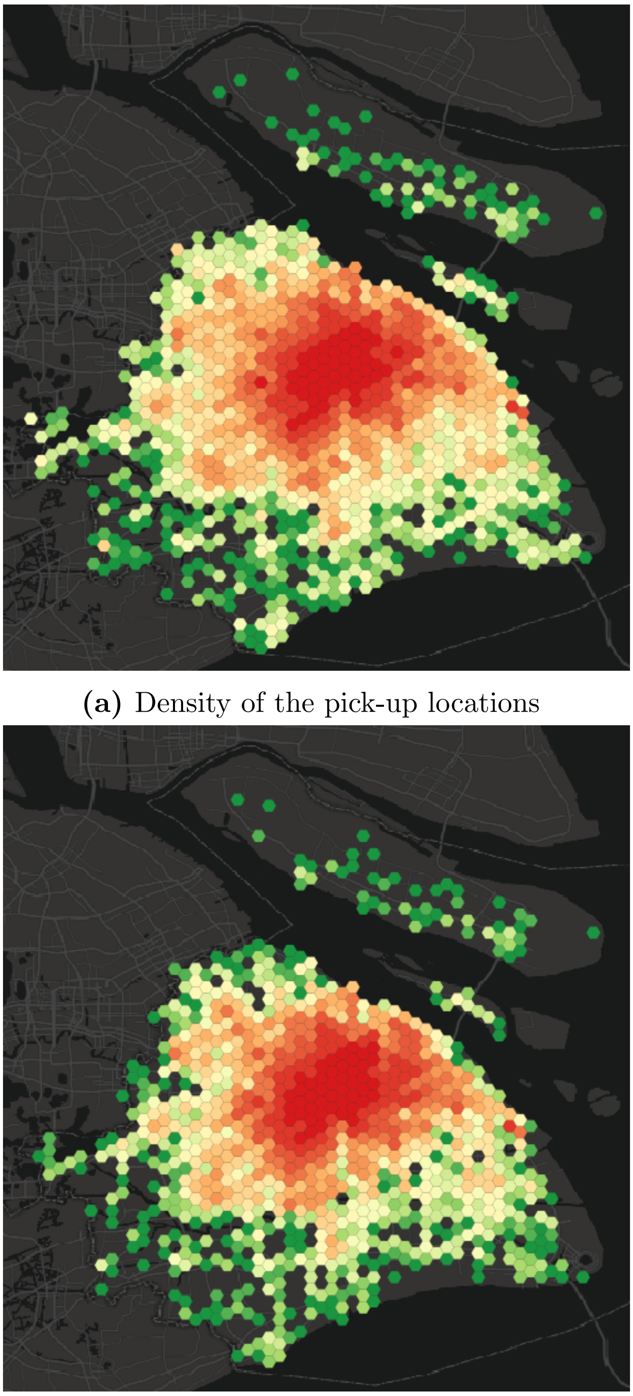
5 Overall patterns

Before going into details of analysis, this chapter summarizes overall patterns from the pre-processing results to give an overview of the taxi service areas, regularity of taxi activities, the distribution of real routes and the shape of the shortest routes.

5.1 Distribution of service areas

The taxi service areas covers large proportion of Shanghai. Figure 5 shows the densities of pick-up and drop-off locations from the recoded FCD data. The more dark the red color is on the map, the higher the density is in this area. It is of no surprise that the largest proportion of both the pick-up and the drop-off locations are centered around Pudong, Xuhui, Huangpu, Yangpu districts, which are the economic and education center of the city. There are some dispersion around the peripheral areas meaning that certain places of interest may attract traffic flows from the center to these places as well as flows back from these areas.

It is also interesting to notice the connection between the main part of the city and the island separated by the river. The density of dots on the island and the other fraction of the city on the other side show frequent commutes between the city center to these places. While it is hard to know exactly what these commuters are based on available data, the spread of taxi activities points to the importance of taxi service in the life of Shanghai.



(a) Density of the pick-up locations

(b) Density of the drop-off locations

Figure 5. Density of the GPS data on pick-up and drop-off locations

5.2 Taxi service activities

This section introduces the taxi drivers' on-service activities across days, the duration of services, and travel distances for different trips.

5.2.1 Regularity of taxi services

Taxi drivers may follow regular routines of their service activities. It is important to know the regularity of taxi services across hours.

Figure 6 summarizes the distribution of pick-up activities across days. The dots in the plot represent the number of pick-up activities at certain hour of the day (1:00 to 24:00), and the lines represents the flow of the time. In general, the figure indicates a strong pattern of taxi drivers' service operations: the most active period is between around 9:00 in the morning and 24:00 in the evening. For the rest of the day, the number of activities drops dramatically and there is less need for taxi services.

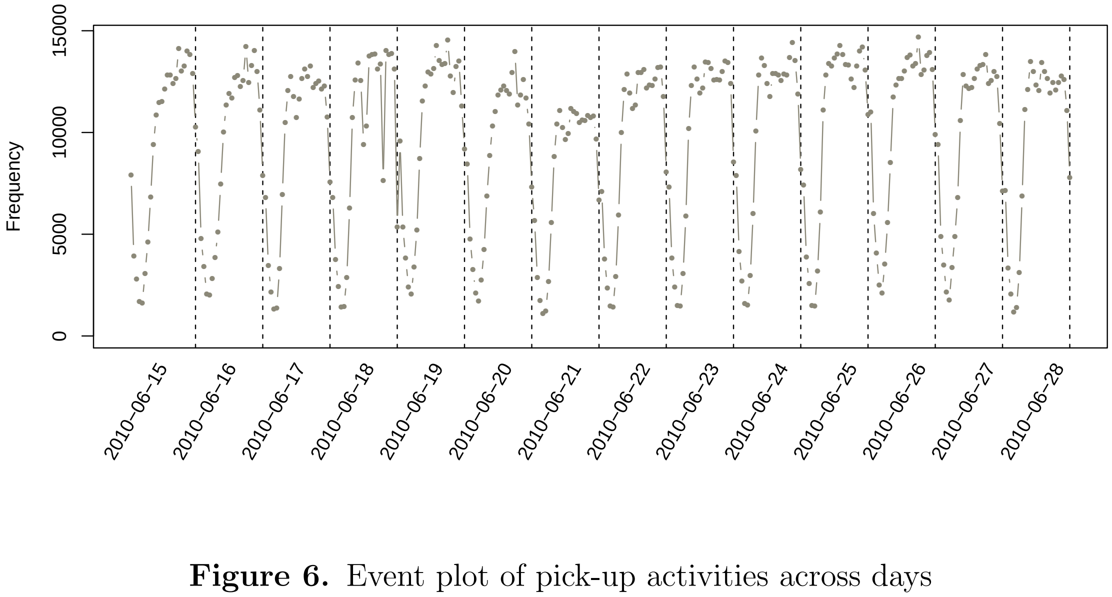


Figure 6. Event plot of pick-up activities across days

Figure 7 illustrates in detail the distribution of taxi pick-up times. It is clear from the figure that the most activities occur between 6:00 and 22:00 when people commute during working hours or experience night life. A small group of people also take taxi at mid-night around 1:00 as the peaked curve shows.

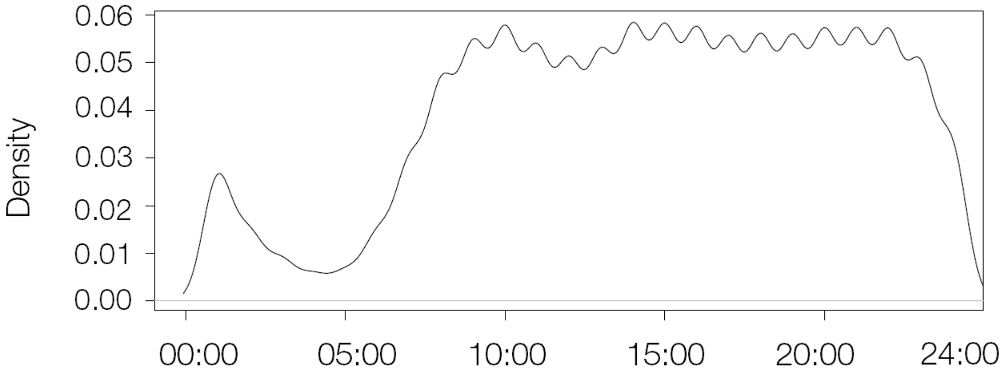


Figure 7. Distribution of pick-up time around the clock

Overall, the taxi services reach its peak in the day time and reduces significantly during mid-night.

5.2.2 Service duration

In addition to pick-up times, the time spent on road also provides important information on regular patterns of taxi services in Shanghai. The average duration is 14.54 minutes with the standard deviation of 16.85 minutes. As Figure 8 shows, most trips take around 0 to 30 minutes while there is variation in different periods of the day (x-axis is the index of travel routes, and the y-axis is the travel duration in minutes). Distributions of service duration over different periods show no differing patterns with each other.

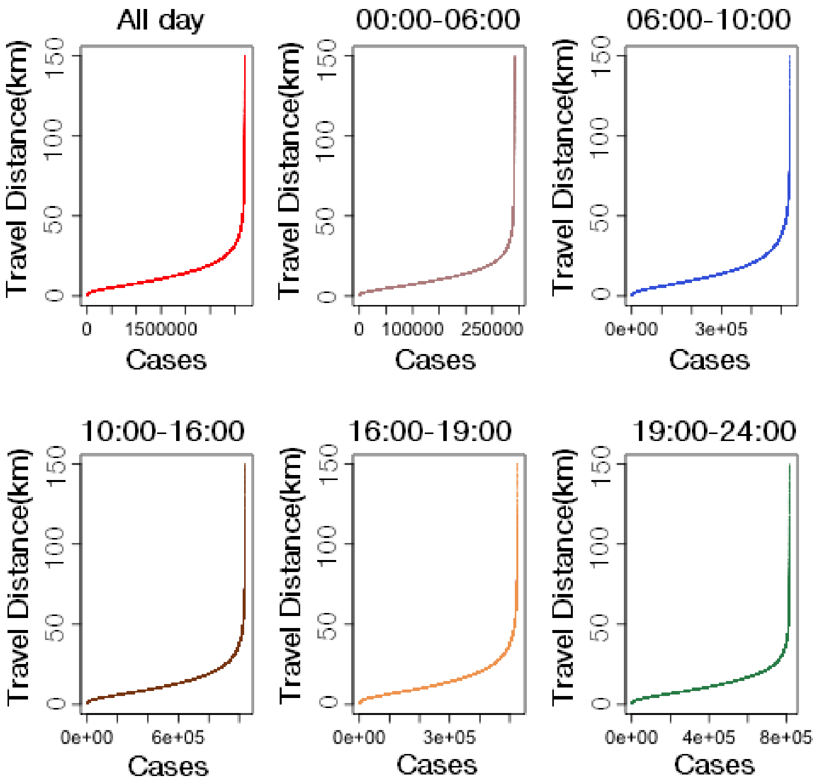


Figure 8. Distribution of travel duration

5.2.3 Travel distances

In addition, Figure 9 shows the distribution of traveled distances of on-service taxis. The average distance is 7.27 km with the standard deviation of 8.05 km. The majority of customers using taxi for short distance traveling. The long right-tail of the density curve till the value around 50 km also means that a small proportion of customers use taxi for long-distance traveling. It shows little variation of travel distances in different hours of the day.

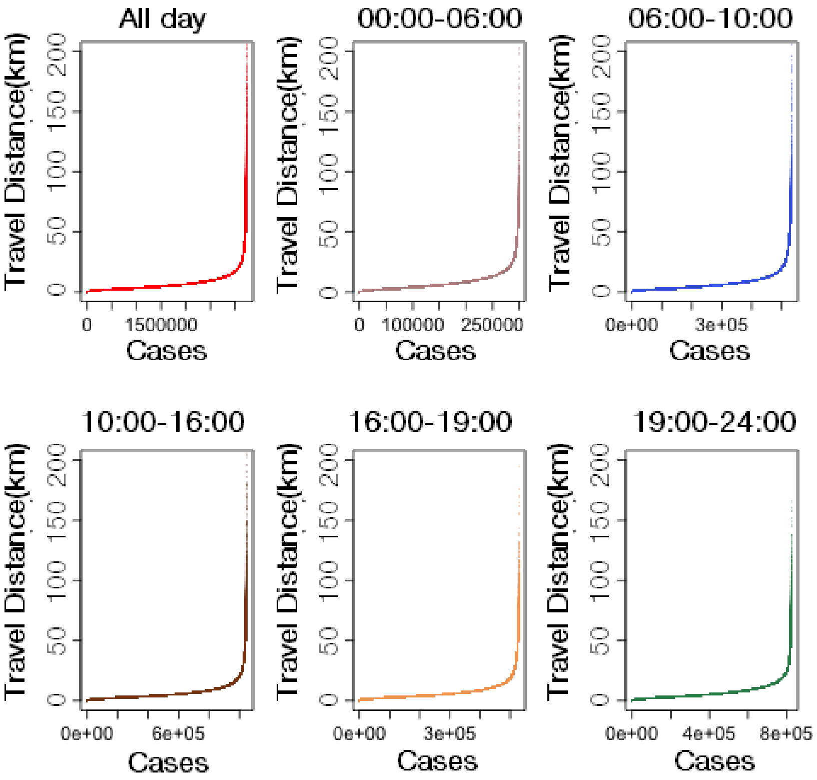
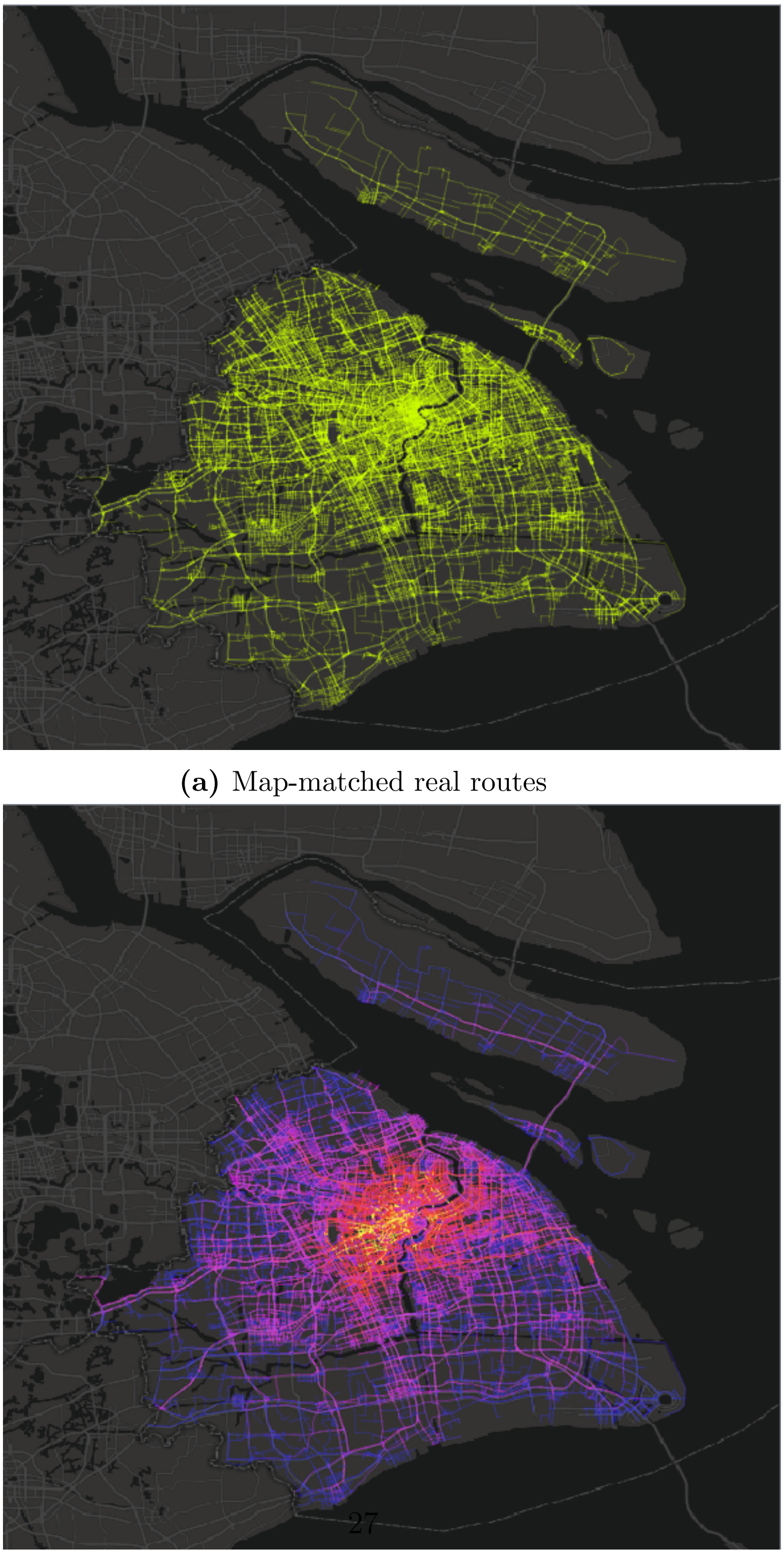


Figure 9. Distribution of travel distance of on-service taxi drivers

5.3 Real routes

Map matching produced satisfying results. To give an overview of map-matched real routes, the following figures show the results of the map matching process.

The left panel of Figure 10 shows the map-matched real taxi routes with passengers on boarder against the Openstreetmap as the source of route networks. To illustrate the usage of different routes, the right panel of Figure 10 in addition visualizes the density of the map-matched routes. The figure shows that while the most frequently-used routes are located around the city center as depicted by yellow and red, there are also some routes travel beyond the city center but used frequently, in particular the route between the city center and the airport.



(a) Map-matched real routes

(b) Density of the routes

Figure 10. Map-matched real taxi routes in Shanghai

5.4 Shortest routes

In addition to the real routes, the shortest routes are also identified and presented. Dijkstra's algorithm was used to calculate the shortest path from the origin to the destination Dijkstra (1959). The algorithm iteratively compares the distance of all the nodes on the paths connecting the origin node and the detonation node, and calculates the shortest route as the collection of nodes that has the shortest distance connecting Origin and Destination (OD).

The following three figures show the examples extracted from the computed dataset. Figure 11 shows a taxi driving route (red line) and its corresponding shortest route (blue line) with relatively short travel distance between 0km and 5km. While there is no overlapping segments between them, the two routes are quite parallel to each other and the taxi driver is likely choose the more straight route.

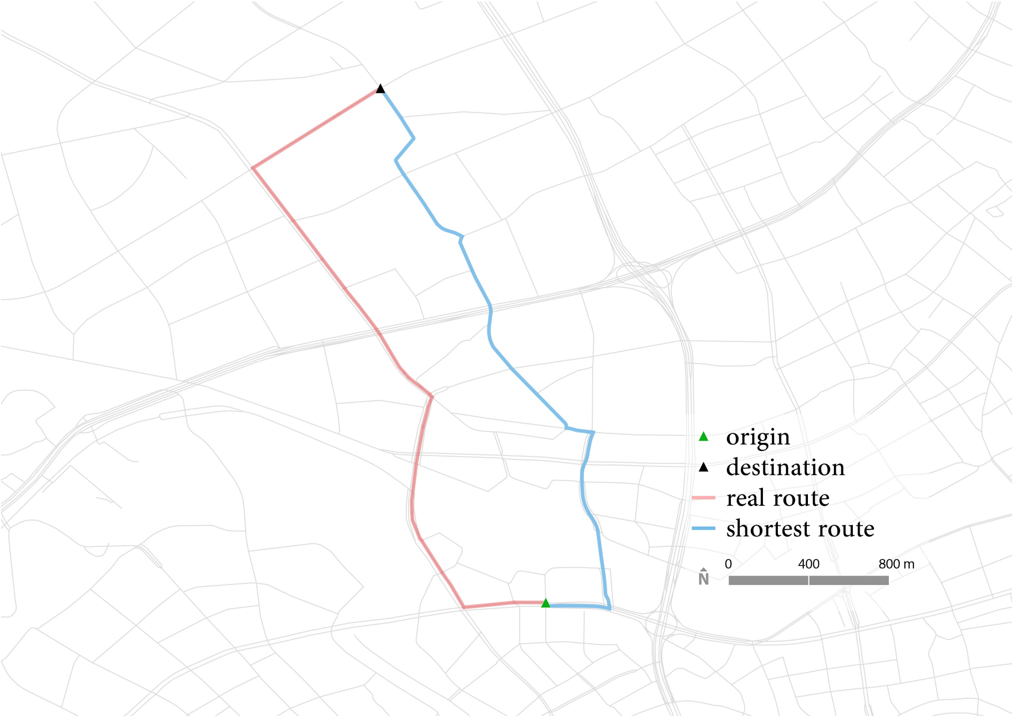


Figure 11. Example 1: the real route and the shortest route

Figure 12 shows a taxi driving route with a long travel distance (longer than 10km) and its computed shortest route. From the origin, two routes diverge with each other, but after around the half of the distance, the real route overlaps with the shortest route. This observed pattern indicates that while other factors may influence the driving behavior of the taxi driver, taxi drivers are more likely to choose the shortest route when possible.

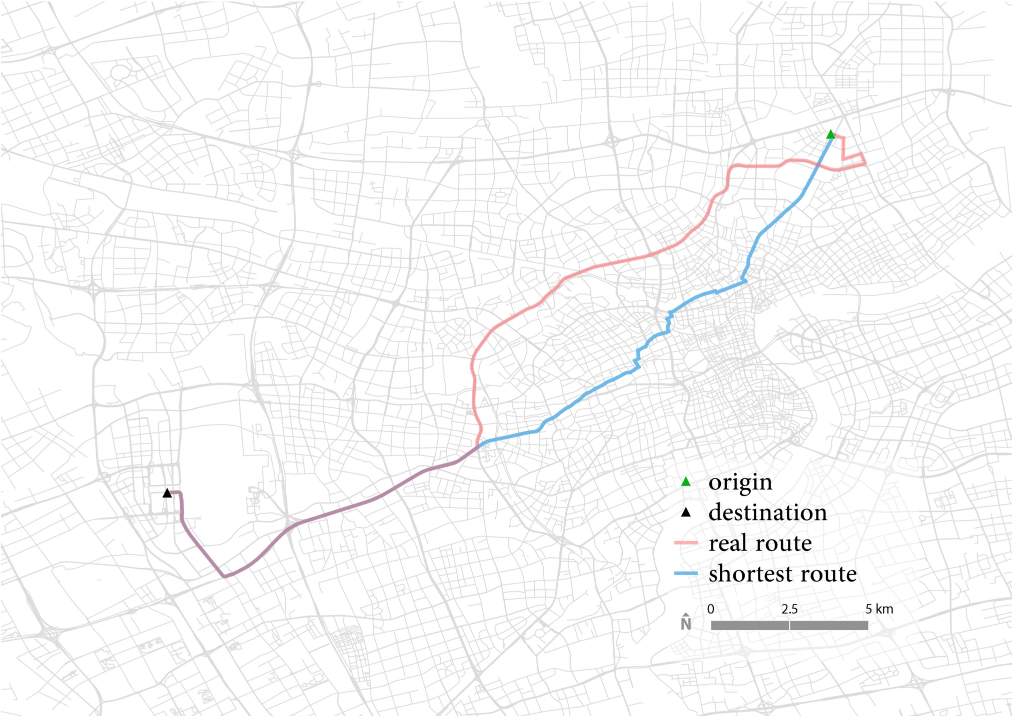


Figure 12. Example 2: the real route and the shortest route

Figure 13 illustrates a route with middle distance (between 5km and 10km) and shows in addition that the travel distance of the shortest route and the real route may not differ so much.

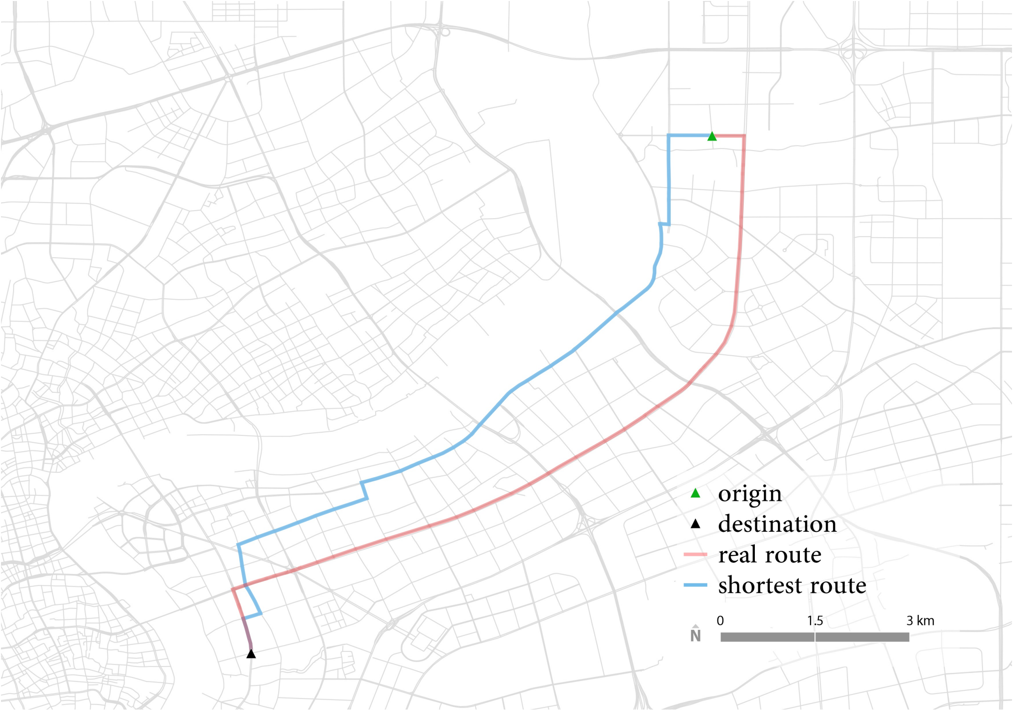


Figure 13. Example 3: the real route and the shortest route

The revealed pattern offers the first impression of both the real and the shortest routes. In general, despite some deviations, the shortest routes have similar distribution with the real routes. However, the extent of match of the shortest routes with the real routes may depends on the travel distances. The next chapter then conducts comparison analysis to find the differences between the real and the shortest routes.

6 Comparison analysis

Comparison analysis is a preliminary step to examine the influence of selected features individually in order to compare the differences between the real and the shortest routes. To compare between real and the shortest routes, a number of elements are considered and are categorized into route-based and segment-based analysis.

To investigate the extent of similarity between the real and the shortest routes, how much longer are the real routes to the shortest routes, and how circuitous the routes are, route-based analysis compares the percentage of shared length (PSL), percentage of length difference (PLD), and route circuity between the real and the shortest routes. Because traffic conditions may differ in different periods of a day and between weekdays and weekends, the variances of PSL, PLD and route circuity are also calculated against different time. In addition, because travel distance may also impact the similarity between the real and the shortest routes due to increasing likelihood of exogenous influences, the PSL, PLD and route circuity are calculated in different groups of travel distances.

In addition to route-based analysis, segment-based analysis takes into account of the features of each segment and how the real and the shortest routes differs in terms of these features. It considers the influence of road type, number of intersections and number of left turns. Table 3 summarizes the variables used in the comparison analysis.

Variables	Route-based analysis	Segment-based analysis
PSL	YES	
PLD	YES	
Route circuity	YES	
Road type		YES
Intersection		YES
Number of left turns		YES

Table 3. Overview of variables in comparison analysis

6.1 Route-based analysis

Route-based analysis consists of comparing the PSL, PLD and route circuity between the shortest and the real routes. The details of the results are introduced in the following subsections.

6.1.1 Percentage of shared length

PSL illustrates the extent to which the real routes are similar to the shortest one. Higher value of PSL indicates that taxi drivers are more likely to follow shortest routes. As aforementioned, due to bounded rationality and impact of a bunch of exogenous factors, the actual routes taxi drivers chosen may deviate from the predicted shortest routes. Taking into account of this fact, calculating the PSL between real routes and the shortest routes is appropriate for analyzing to what extent the real routes follow the prediction of the shortest routes.

The PSL is calculated by first identifying the overlapped segments of real routes and the shortest routes and then dividing the overlapped length by real total length. In formula,

$$\mathrm{PSL} = \frac{\sum_{i \in \Phi} r_i}{\sum_{j \in \Theta} r_j}$$

where Φ and Θ denote the set of overlapped segments and the segments contained in real routes, i denotes individual segments, and r_i denotes the segment length.

6.1.1.1 Overall result

Overall, it is expected that the real routes deviates from the shortest but not too much. Table 4 summarizes the distribution of PSL from the whole data. On average, there are around 46% of route segments that are overlapped with the shortest routes. The completely matched routes account for 16.7% of the total routes. There are also poor matches that have PSL values lower than 20%, which account for 31.4% of the total routes. There are 41.9% of routes that have more than 50% overlap with the their corresponding shortest routes.

PSL	Percentage (%)	Cumulative percentage(%)
1	16.66727	16.66727
0.8 -1	5.54906	22.21633
0.5 - 0.8	19.64954	41.86587
0.2 - 0.5	26.71682	68.58269
0 -0.2	31.41731	100
Average:	0.4565	

Table 4. Distribution of PSL

6.1.1.2 Variance between different time periods

The PSL values may differ between different time periods because during rush hours the taxi drivers may prefer to choose relatively longer routes to avoid congestion. And the traffic situation between weekdays an weekends may also differ.

However, from the Figure 14 which shows the distribution of PSL grouped by different time slots in a day (0:00-6:00, 6:00-10:00, 10:00-16:00, 16:00-19:00, 19:00-24:00), the PSL values do not vary much over different time periods .

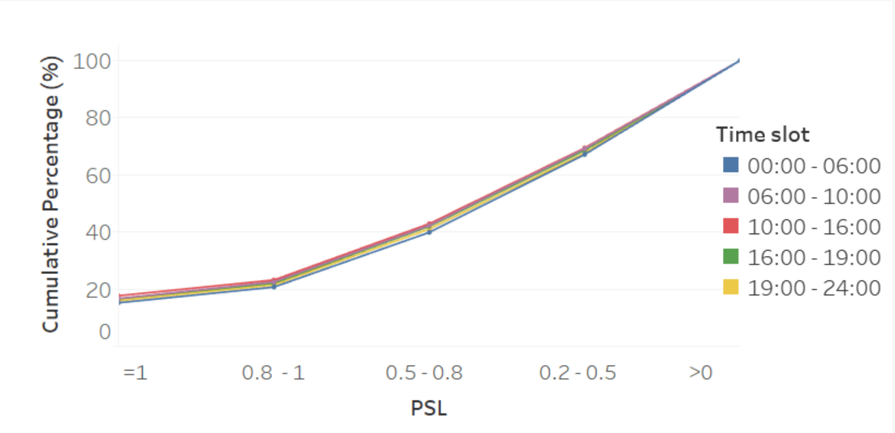


Figure 14. PSL grouped by time slots

In addition, Figure 15 distinguishes the distribution of PSL between weekdays and weekends. Again, the figure shows no distinguishable patterns as two cumulative percentage lines almost overlap with each other.

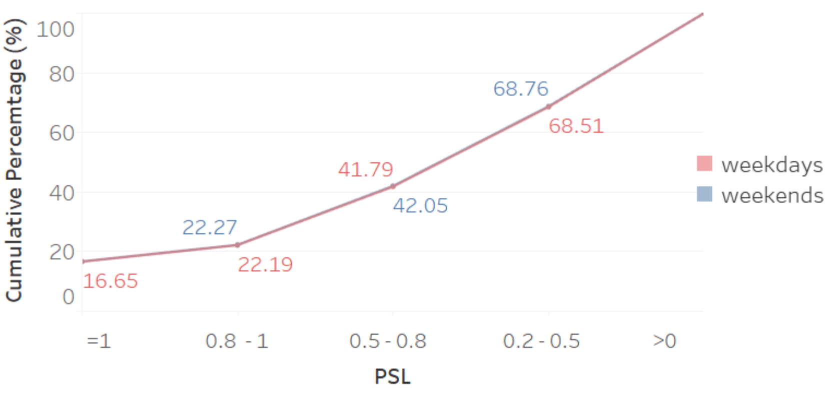


Figure 15. PSL grouped by weekdays and weekends

While time impacts little on the difference of PSL between the real and the shortest routes, travel distance may have impacts on the PSL values. The following analysis introduces the results by comparing between different travel distances.

6.1.1.3 Variance between different travel distances

When the PSL is grouped by the distances of real routes, as shown in Figure 16, it is clear that with the increase of route length, taxi drivers tends to deviate more from the shortest path. For trips less than 5 km as shown in blue line, more than 50% of routes have PSL values larger than 0.5; but for trips longer than 10 km, the percentage decreases to less than 20%. One of the possible reasons is that shortest travel distances would reduce the number of possible connections between the OD.

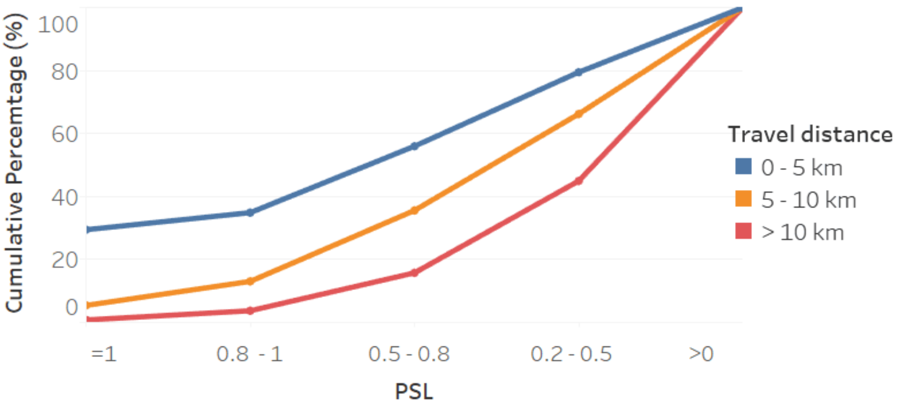


Figure 16. PSL grouped by lengths of real routes

And indeed, the correlation between OD distance and the PSL shows a negative value of -0.32, which was calculated by $r_{xy} = \frac{s_{xy}}{s_x s_y}$, where s_{xy} is the sample covariance and s_x and s_y are sample standard deviations. Therefore, the longer the travel distances, the more diverse are the real routes compared to the shortest routes.

6.1.2 Percentage of length difference

Because the lengths of different segments differ and PSL only calculates the overlapped lengths, it is also important to know the length difference between the real and the shortest routes. PLD compares the total length between the real and the shortest routes to show how much longer is the actual route to the shortest routes. In formula,

$$\text{PLD} = \frac{\sum_{i\in\Gamma} r_i - \sum_{j\in\Delta} r_j}{\sum_{i\in\Gamma} r_i}$$

where Γ and Δ denote respectively the set of real route segments and the shortest route segments, i denotes an individual segment, and r_i and r_j denotes the length of the segment. Or in short: PLD $= \frac{L1-L2}{L1}$, which means the value is calculated by subtracting the length of the shortest route ($L2$) from the real route ($L1$) and then divided by the length of the real route.

6.1.2.1 Overall result

Even though real routes deviate from the shortest routes to some extent, it is unlikely that the real routes have a very large proportion of extra length compared to the shortest routes. From the distribution of PLD values as shown in Table 5. The mean value of PLD is around 16%. The majority of the routes have PLD values lie between 0 to 0.2, meaning that most of real routes are only slightly longer than the shortest routes. Routes with PLD values larger than 0.5 only account for less than 5% of total routes. Overall, the majority of real routes has a mild addition of distances to the shortest routes.

PLD	Percentage (%)	Cumulative percentage(%)
1	0	0
0.8 -1	0.6421462	0.6421462
0.5 - 0.8	4.2816988	4.923845
0.2 - 0.5	26.540425	31.46427
0 -0.2	68.53573	100
Average:	0.1609	

Table 5. Distribution of PLD

6.1.2.2 Variance between different time periods

Similar to PSL, it is possible that the values of PLD differs between different time periods. Figure 17 shows that the distribution of PLD across hours with different lines representing different time slots. Again, as the lines largely overlap with each other, the values of PLD are quite similar in different periods of a day.

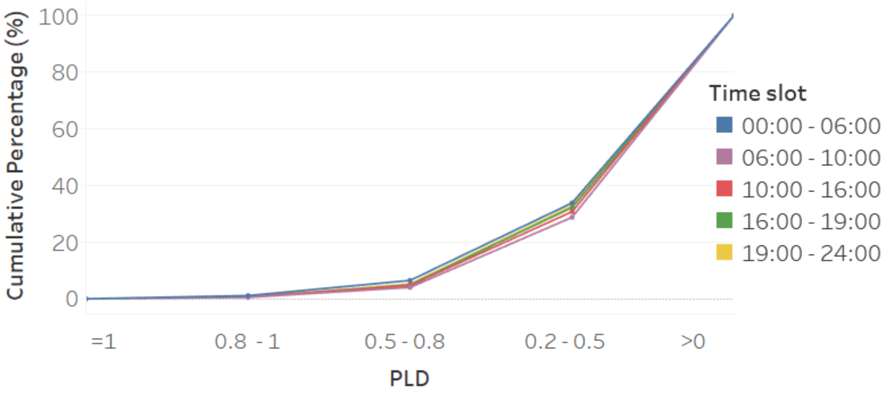


Figure 17. PLD grouped by time slots

Like PSL, Figure 18 shows that PLD is similar between weekdays and weekends.

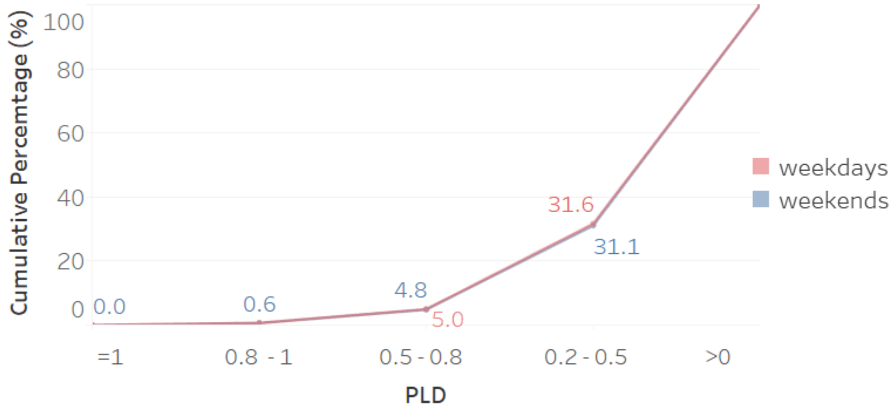


Figure 18. PLD grouped by weekdays and weekends

Therefore, time also impacts little on the difference of PLD between the real and the shortest routes.

6.1.2.3 Variance between different travel lengths

Between different travel distances, the distribution of PLD differs. As Figure 19 shows, when the travel distances increase, real routes grow longer than the shortest routes. As shown with the red line, with travel distances larger than 10 km, the routes with PLD values larger than 0.2 account for around 50% of total routes. When the trip distances are smaller (less than 5 km), the routes with PLD values larger than 0.2 only represent around 20% of total routes (blue line).

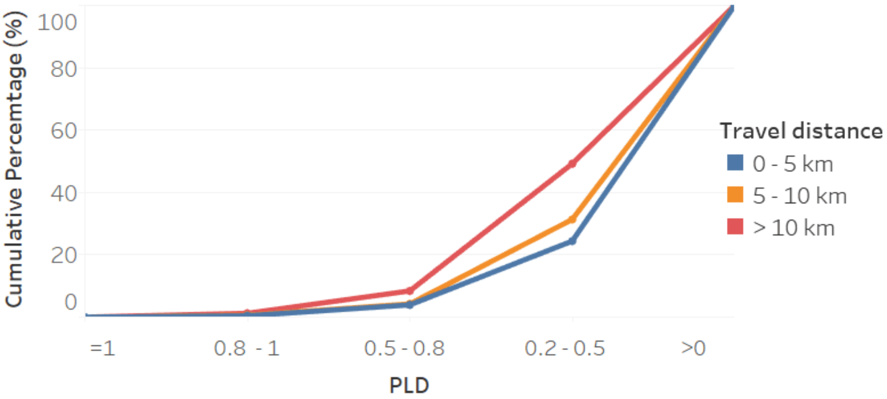


Figure 19. PLD grouped by lengths of real routes

From the above comparisons, it is clear that the real routes have some deviation to the shortest routes, and the real routes are relatively longer. The deviation between the real and the shortest routes differs in terms of travel distances but not so much between different time periods.

6.1.3 Route circuity

Route circuity is also an important feature to compare between the real and the shortest routes. In essence, drivers might prefer straight routes than circuitous routes.

Route Directness Index (RDI) was generated to measure route circuities and compare them between the real and the shortest routes (Papinski and Scott, 2011).

For each OD pair a, b, the route circuity is calculated as follows,

$$rcr = \frac{l_{ab} - d_{ab}}{l_{ab}}$$

, where l_{ab} refers to the route distance, and d_{ab} refers to the geometric distance between point a and point b, and

$$d_{ab} = \arccos(\sin\theta_a * \sin\theta_b + \cos\theta_a * \cos\theta_b * \cos\Delta\lambda) * R$$

where θ refers to latitude, λ refers to longitude, R refers to the radius of the earth.

The range of circuity values is between 0 and 1. The closer the value is to 0, the more straight is the route; and the closer the value is to 1, the more circuitous is the route.

6.1.3.1 Overall results

Overall, the shortest routes have lower circuity than the real routes. The average value of circuity of the real routes is about 0.32 but the value for the shortest routes is only 0.19. The majority of the real routes have circuity values between 0.2 and 0.5 while the circuities of the majority of the shortest routes lie between 0 and 0.2. The circuities of both the real and the shortest routes are not equal to 0. The results make sense because the shortest routes tend to choose short and straight route segments than circuitous segments.

Route Circuity	Percentage in real routes (%)	Cumulative percentage(%)	Percentage in shortest routes (%)	Cumulative percentage(%)
1	0	0	0	0
0.8 -1	1.1759	1.1759	0.09560	0.0956
0.5 - 0.8	12.1467	13.3225	1.3021	1.3977
0.2 - 0.5	62.7011	76.0236	37.6061	39.0037
0 -0.2	23.9764	100	60.9963	100
Average	Real routes:	0.3154		
	Shortest routes:	0.1862		

Table 6. Distribution of Route Circuity of real routes and shortest routes

6.1.3.2 Variance between different time periods

Because time may also impact taxi drivers' choices, it is also helpful to compare the route circuity between different time periods. Similar to the results from PSL and PLD, the route circuity differ little across time slots. As shown in Figure 20, lines in different colors representing different time periods mostly overlap with each other.

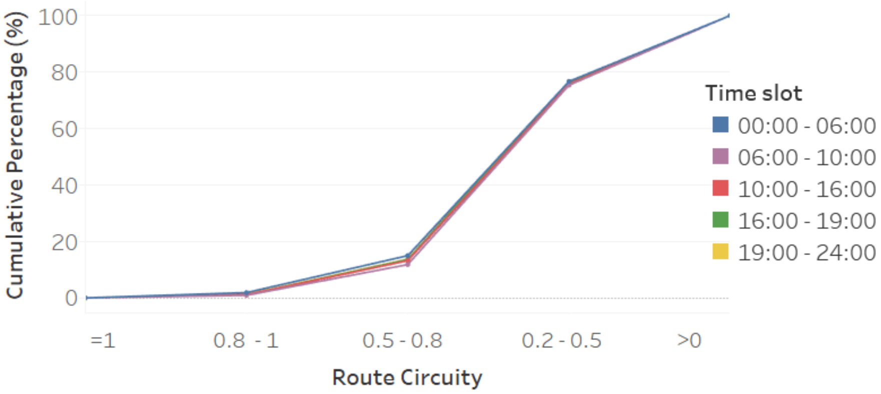


Figure 20. Route circuity grouped by time slots

The same is true for comparing between weekdays and weekends. As shown in Figure 21, the difference between the route circuity during weekdays and weekends are ignorable.

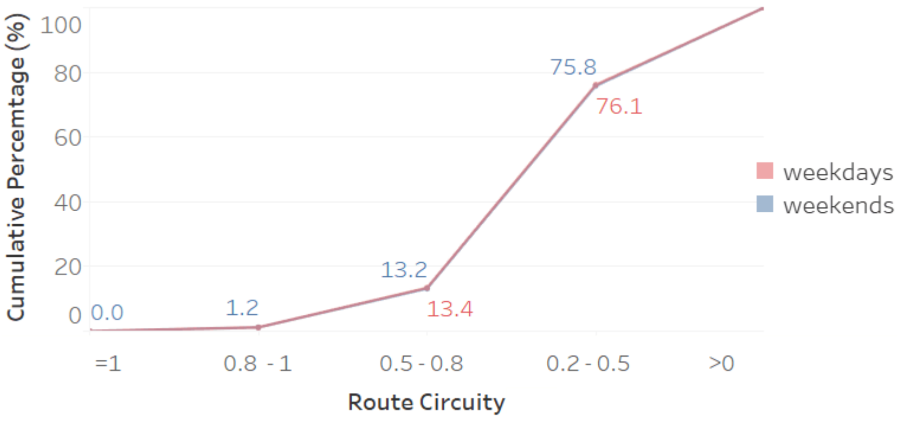


Figure 21. Route circuity grouped by weekdays and weekends

In sum, the difference between the real and the shortest routes is insignificant across different time periods.

6.1.3.3 Variance between different travel distances

As mentioned before, circuity values of different routes are also compare across travel distances. Figure 22 shows that the longer the travel distances, the more circuitous are the routes. Around 90% of routes have circuity values larger than 0.2 when the travel distances are larger than 10 km. When the travel distances are short, that is less than 5 km, only about 70% of routes have circuity values larger than 0.2. It is therefore clear that the travel distances significantly influence route circuity.

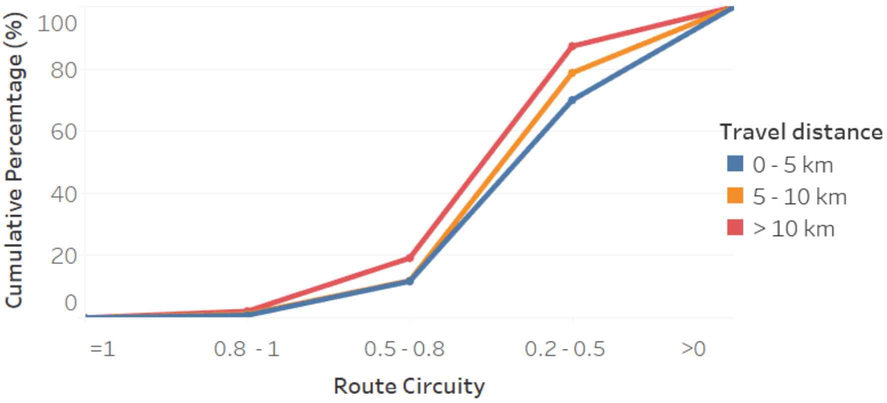


Figure 22. Route circuity grouped by length of real routes

The above comparison analysis on route-based features show the derivation of the real routes from the shortest routes. The difference between the real and the shortest routes becomes more significant if the travel distances are taken into account. Overall, most real routes have overlaps with the shortest routes (46% on average) and have larger length (16% on average), and they are more circuitous than the shortest routes.

6.2 Route segment-based analysis

Segment-based analysis will examine the effects of road type, number of intersections and turns on the taxi drivers' route choice.

6.2.1 Road type

A complete travel route consists of multiple segments. The combination of segments in a chosen route also reveals the preference of the driver. In particular, the road type of each segment may influence the choice of taxi drivers. Thus, percentage of each road type in each route was computed to find the most preferable type of roads by taxi drivers.

Road types play important roles in influencing drivers' route choice behavior. Because different types of roads have different speed limits, individual drivers may prefer to choose high level roads with high speed limit to shorten travel time. For taxi drivers, however, their route choices are also influenced by customer demands which may reduce the flexibility in choosing the type of roads that have the highest speed limit. From the taxi driver data in Shanghai, it is able to derive and compare features of the original chosen routes and the shortest routes in terms of road type.

As traditionally categorized, five types of road based on Open Street Map data are analyzed: motorway, primary, secondary, tertiary and residential roads. Motorway has the highest speed limit up to 120 kilometers pro hour in Shanghai. Primary, secondary and tertiary roads respectively have speed limits of 70, 60 and 40 kilometers pro hour. Residential roads are located within residential areas and have a speed limit of 40 kilometers pro hour as the same with the tertiary roads.

Figure 23 shows the share of different types of roads in both the real and the shortest routes. In real routes, primary ways take the largest proportion while the secondary ways have the second largest proportion. And the resiential roads have the smallest share compared to other four types of roads. However, in the shortest routes, the secondary roads have the highest share and the primary ways only have the second largest share. And the motorway in the real routes have the smallest proportion compared to other types of roads.

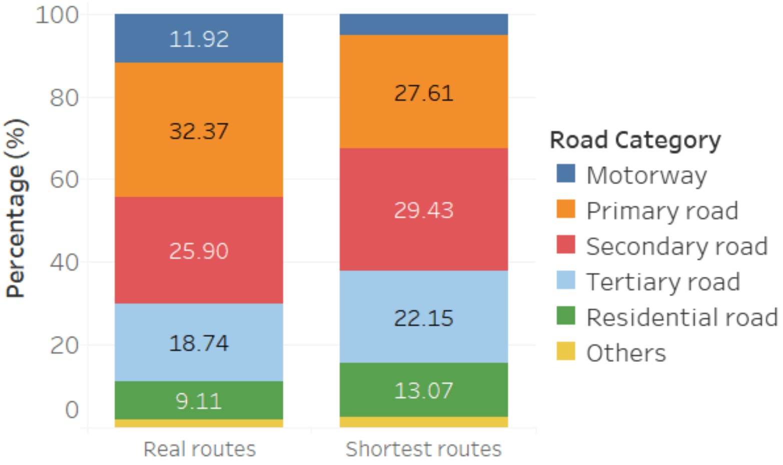


Figure 23. Share of different types of roads in the real and shortest routes

In addition to average values, Figure 24 show the percentage usages of different types of roads in different travel distances. On average, taxi drivers prefer primary roads more than other types of roads. Because primary roads have relatively high speed limit and can accommodate huge flows of traffics, and thus they are a major type of roads in urban areas. However, the usage of primary roads differs across travel distances. In short-distance trips that shorter than 5 km, the primary roads share lower percentage compared to the mid- and long-distance trips. That holds true for both the real and the shortest routes.

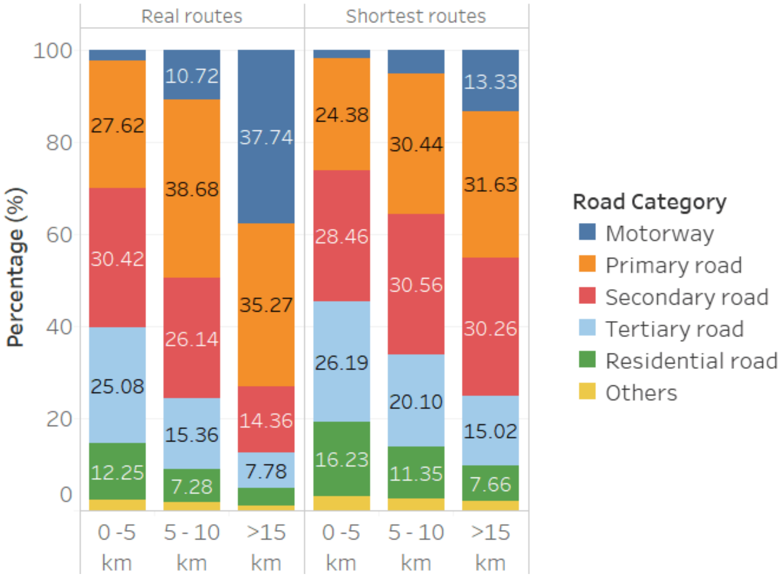


Figure 24. Share of different types of roads across travel distances

Even though motorways have highest speed limits, they are on average less used by taxi drivers because most trips are short distant. It is of no surprise because most of the trips are around the city center and there is no need to use motorways which are usually used to travel across or outside the city. And indeed, in long trips that are larger than 10 km, the usage of motorway increases significantly to 37% as shown in the figure.

On average, the secondary ways have the second highest usage in real trips. However, compared to primary and secondary roads, percent usage of tertiary roads in original routes are lower. This result follows the logic that taxi drivers are more likely choose the roads with high speed limit than those with low limit.

The result also shows that compared to the shortest routes, the real routes have higher share of the primary roads and the motorways. This is an indicator showing that taxi drivers have preferences for the types of roads that have high speed limit.

Figure 25 also shows the share of different types of roads in different time periods within a day. Actually, time difference on shares of different road types is hardly distinguishable. Across hours, the primary ways have the highest share while the residential roads have the lowest. But in the shortest routes, the secondary ways instead have the highest share.

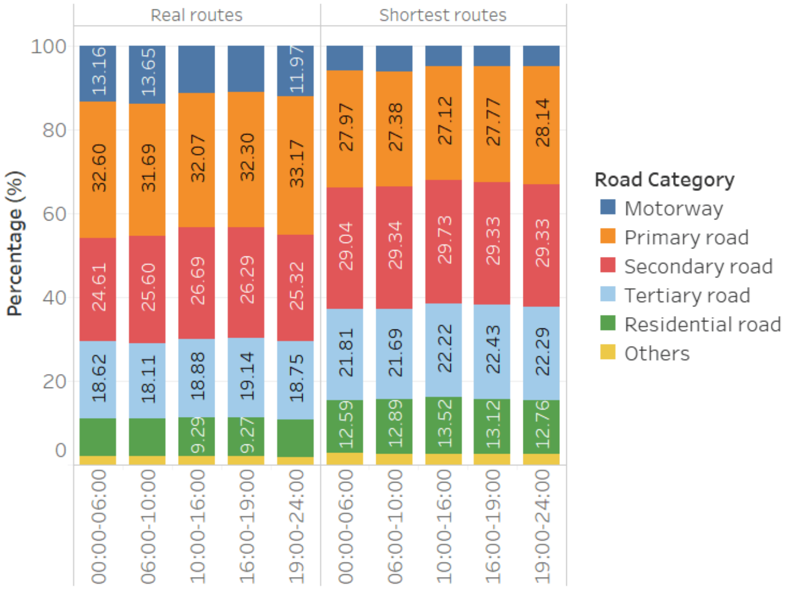


Figure 25. Share of different types of roads across time slots

Figure 26 show in addition the share of different road types in both weekdays and weekends. Again the difference is ignorable between the weekdays and the weekends. The same results hold for the share of different types of roads in both the real and the shortest routes.

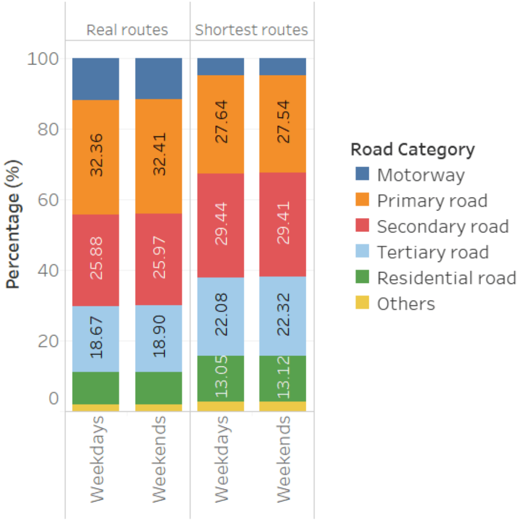


Figure 26. Share of different types of roads in weekdays and weekends

In a nutshell, the difference between the real and the shortest routes in terms of road types reveals the preferences of taxi drivers. They prefer to choose the roads with high speed limit instead of choosing directly the those with the shortest connection. In particular, when the travel distances increase, taxi drivers prefer more the motorways that have the highest speed limit.

6.2.2 Intersections and number of turns

Besides road types, other features like (complex) intersections and number of turns may also influence the choice of taxi drivers. The following results shows the difference between the real and the shortest routes in terms of these features.

As shown in Figure 27, over different travel distances, the distributions of numbers of straight ways between the real and the shortest routes are quite similar [2]. Thus, straight ways may not be an important consideration for taxi drivers.

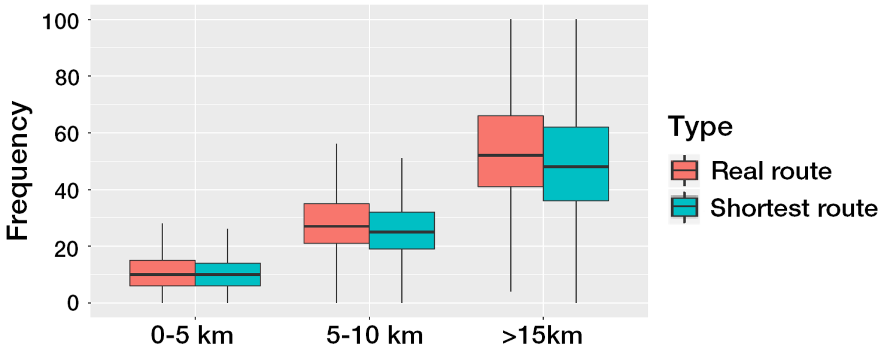


Figure 27. Comparison of numbers of straight ways in the real and the shortest routes

Because Chinese regulations requires right-hand traffic, drivers are also commonly assumed to avoid left turns (Yao et al., 2013). However, Figure 28 shows that in comparison to shortest routes, taxi drivers actually chose the routes with relatively more left turns. When even taking into account of the number of sharp left turns as shown in Figure 29, real routes have more sharp left turns on average than the shortest routes [3].

[2] The straight ways are the ways whose nodes are nether the left- nor the right-turns, or the ways that only connect with two nodes. The left turns are identified when the degree of direction is between 210 and 330; and the right turns are identified when the degree of direction is between 30 and 150.

[3] Sharp left turns are identified when the degree of direction is between 210 and 270

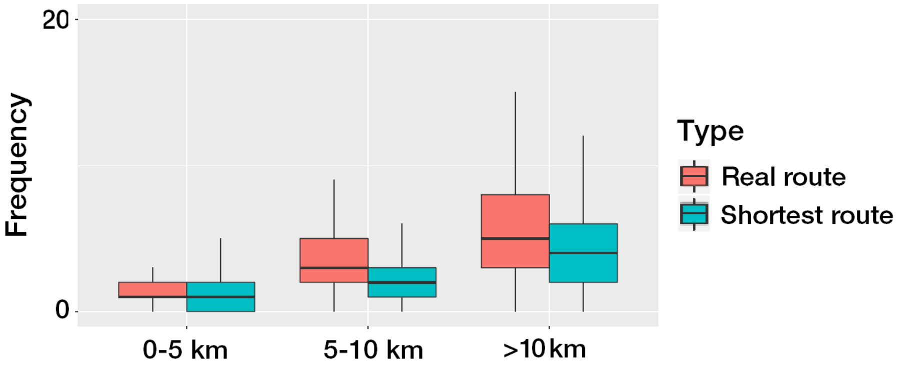


Figure 28. Comparison of numbers of left turns in the real and the shortest routes

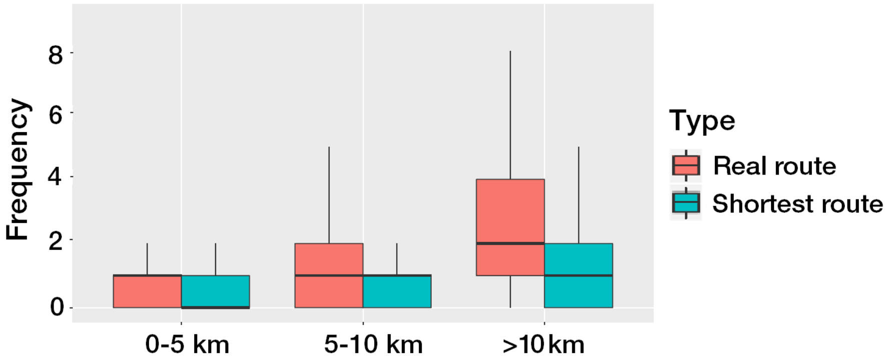


Figure 29. Comparison of numbers of sharp left turns in the real and the shortest routes

Finally, the impact of number of intersections on taxi drivers' route choices was considered. Figure 30 shows that real routes and the shortest routes are actually very similar in terms of number of intersections. In other words, taxi drivers will not deliberately choose a longer route in order to avoid intersections.

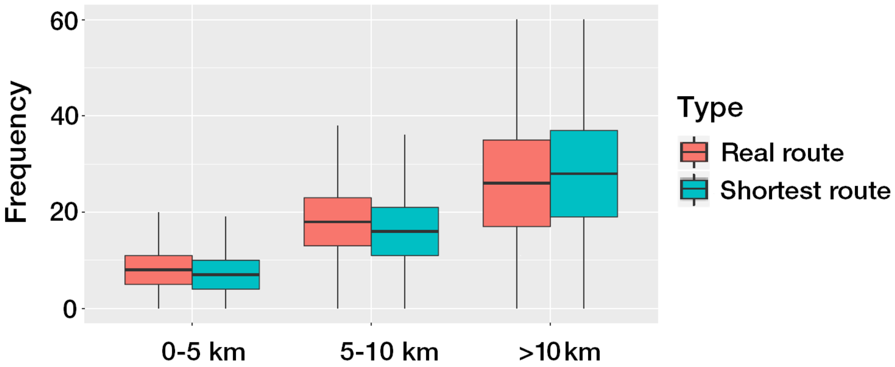


Figure 30. Comparison of numbers of intersections in the real and the shortest routes

In sum, the above analysis compare both the route- and the segment-based factors in influencing taxi drivers' choices. In expectation, real routes deviate from the shortest routes and are longer than shortest routes by 60% on average. The longer taxi drivers travel, the higher the deviation of real routes from the shortest routes. Time periods and whether the travel date is in weekdays or on weekends seem to influence little on the route choices. In contrast, taxi drivers prefer certain road types than others. In particular, the primary roads are the most frequently used type of roads compared to secondary, tertiary, residential and motorways. But the comparison analysis show that almost no influence in terms of straight ways, left turns, sharp left turns and intersection exist when comparing real routes and the shortest routes.

7 Preference analysis

Comparison analysis considers only a single factor once at a time as described in the previous chapter. And insignificance of some factors may due to the existence of interaction effects that is hard to be detected with simple comparison. Therefore, a more sophisticated way to analyze the preferences of taxi drivers need to be conducted.

The most thorough way to compare multiple factors is to identify all alternative routes and find the factors of significance by conducting a regression analysis. However, calculating all alternative routes and compare their relevant features is computationally demanding. To get rid of complexity to find all alternative routes in a network, the K shortest path algorithm was introduced by imposing constrains on route choices and limiting the number of alternative routes.

The idea originates from the seminar paper by Dijkstra (1959) in which he proposed the problem of constructing the shortest path from n node. Later on, the algorithm was developed further to allow the calculation of multiple routes with a given set of ODs (Hoffman and Pavley, 1959; Ishii, 1978; Yen, 1971). The core of the later algorithm lies in the construction of scenarios or weights by which the routes are chosen to minimize the weighted costs. From identifying the shortest routes with specified weights, the resulting analysis is able to compare between different scenarios and find the order of factors by their influence on the final choices.

7.1 Scenarios and their application to general routes

In this study, 10 scenarios have been specified to detect the extent of influence by multiple factors. The basic scenario (scenario 0) contains only the weights of travel distances. And the resulting routes correspond to the shortest routes by length as previously discussed. The rest of the scenarios are as follows.

- Scenario 1: The first scenario in comparison adds traffic free flow to the weights to account for the influence of travel time. The free flow is calculated as the shortest possible time t_i at the maximum speed of road segment i, assuming that there is no congestion in the whole traffic network. C_1 is used as the cost to calculate the fastest route in this scenario.

$$C_1 = t_i$$

- Scenario 2: Due to the fear of congestion and delay, taxi drivers may prefer the route with the time-dependent travel time. The time-dependent travel time in each road segment i is calculated as the average travel time T_i by all on-service taxis passing through that segment i of a single day. The travel time is divided into five time period α as mentioned in the previous sections, from 00:00-06:00, 06:00-10:00, 10:00-16:00, 16:00-19:00, 19:00-24:00. And the route with the lowest cumulative observed travel time is identified as the least-observed travel-time route. C_2 is used as the cost to calculate the weighted fastest route in this scenario.

$$C_2 = T_{i|\alpha}$$

- Scenario 3: The third scenario combines travel distance with road types. It corresponds to the route choice behavior that mainly concerns about shortest and high quality routes. In this case, l_i represents the length of road segment i, factors w_i are assigned based on the category of each road segment i. "Motorways" are given a weight of 1; "Primary roads" are given a weight of 2; "Secondary roads" are given a weight of 3; "Tertiary roads" are given a weight of 4; "Residential roads" are given a weight of 5. C_3 is used as the cost to calculate the weighted shortest route in this scenario.

$$C_3 = l_i * w_i$$

- Scenario 4: The fourth scenario takes the effect of free flow and road types into account. C_4 is used as the cost to calculate the weighted fastest route in this scenario.

$$C_4 = t_i * w_i$$

- Scenario 5: The fifth scenario combines travel distances with road usage frequency as suggested by Dijkstra (1959). In this scenario, road usage frequency f_i is based on the frequency of road segment i used from all observed taxi trips during the whole study period, the 14 consecutive days. Because road usage frequency has high standard deviations, the calculation may suffer from numerical instability. Thus, the logged usage frequency is used instead of the original values. C_5 is used as the cost to calculate the weighted shortest route in this scenario.

$$C_5 = l_i \ / \ log(f_i)$$

- Scenario 6: The sixth scenario considers the effect of free flow in addition to road usage frequency, which can be obtained the same in Scenario 5. C_6 is used as the cost to calculate the weighted fastest route in this scenario.

$$C_6 = t_i \ / \ log(f_i)$$

- Scenario 7: In addition to travel distances, Scenario 7 also considers the effects of numbers of left turns. Due to the right hand rule in the regulation of China's traffic, taking left turns means more waiting time in face of traffic lights, and might be more time consuming. By using K shortest path algorithm, for each OD pairs, trips with top ten shortest distances are generated, then the numbers of left turns are calculated for each trip. Trips with minimal number of left turns are chosen as the final results in this scenario. C_7 is used as the criteria to calculate the weighted shortest route in this scenario.

$$C_7 : l_i \ \ considering \ \ min(N_{leftturns})$$

- Scenario 8: Similar to Scenario 7, trips with top ten minimal time are generated, and for each OD pairs, trips with minimal number of left turns are chosen as the final results in this scenario. C_8 is used as the criteria to calculate the weighted fastest route in this scenario.

$$C_8 : t_i \ \ considering \ \ min(N_{leftturns})$$

- Scenario 9, 10: In addition to left turns, Scenario 9 and 10 considers that taxi drivers might prefer taking straight roads instead of taking turns. Trips with top ten shortest distance of minimal time are generated in these two scenarios. And the sum of numbers of left turns or right turns of each trips are calculated from these trips. Trips with minimal number of all turns are chosen as the final results, which are also seen as the most straight ways. C_9 and C_{10} are used as the criteria to calculate the weighted route.

$$C_9 : l_i \ \ considering \ \ min(N_{turns})$$

$$C_{10} : t_i \ \ considering \ \ min(N_{turns})$$

Table 7 summarizes the factors used in each scenarios.

	Scenarios										
Factors	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Length	■			■		■		■		■	
Free flow		■			■		■		■		■
Observed travel time			■								
Road type				■	■						
Road use frequency						■	■				
Left turns								■	■		
Turns										■	■

Table 7. Scenarios: the gray cells indicate the inclusion of the factor into the corresponding scenario.

To calculate the prediction accuracy of different scenarios and thereby determine the extent of influence by varying factors, 10000 routes were sampled from the original data, and PSL was computed against real routes. Table 8 summarizes the results from the predictions of different scenarios.

Scenario	Weighted Criteria	Mean(%)	Over 75% match	Improvement
0	Minimal distance	46.61	25.69	
1	Least freeflow travel time	51.30	30.77	YES
2	Least observed travel time	51.75	30.92	YES
3	Minimal (road category * distance)	48.79	29.39	YES
4	Minimal (road category * time)	45.80	26.72	NO
5	Minimal (road usage frequency * distance)	53.56	32.52	NO
6	Minimal (road usage frequency * time)	52.48	31.94	YES
7	Minimal (left turns * distance)	46.61	25.60	NO
8	Minimal (left turns * time)	51.06	30.50	YES
9	Minimal (turns * distance)	36.23	14.52	NO
10	Minimal (turns * time)	38.46	16.53	NO

Table 8. Prediction accuracy of different scenarios (cells show the cumulative percentage of cases within specified accuracy category)

Scenario 1, 2, 3, 5, 6 and 8 show improvement over the basic travel-distance scenario (scenario 0). In particular, scenario 5 which considers the effects of travel distances and logged usage frequency has the highest prediction accuracy. This means that taxi drivers prefer to choose the routes that are used more frequently by other taxi drivers.

Scenario 1 consider the impact of the speed limits on the taxi drivers' route choices. It shows that taxi drivers prefer roads with high speed limits and lower free-flow travel time compared to minimal-distance routes. The least observed travel time in Scenario 2 was calculated based the average speed of all taxi drivers on each route segment. The improvement of prediction in this scenario indicates that taxi drivers care about the real travel time in choosing the routes.

Scenario 3 shows the scenario in combination with travel distance and road category. The improvement over the basic scenario indicates that preference over road types are also important in addition to real travel distances. Scenario 4 shows that the impact of road category and freeflow travel time together is almost identical to the impact of travel distances. That means the travel distance is still a major concern for taxi drivers' route choice.

Scenario 5 and 6 consider the impact of logged road usage frequency on route choices. Both considering in addition the distance (Scenario 5) and the time (Scenario 6), the road usage frequency contributes to the improvement of prediction accuracy on the real routes.

In Scenario 7 and 8, the number of left turns are considered. The prediction routes are identified by minimize the number of left turns given a set of shortest routes (Scenario 7) or fastest routes (Scenario 8). The results show that the number of left turns has little effect in both situations considering either travel distance or travel time.

The last two scenarios—Scenario 9 and 10—consider the effect of the number of turns. However, the results show very poor prediction from the number of turns. That means taxi drivers consider little about the number of turns in choosing their routes.

To sum up, taxi drivers in general concern about travel distance, travel time, free flow, road type, the number of left turns and road usage frequency. However, taxi drivers care little about the number of turns. Overall, the travel distance has the highest prediction power which means that it is still the most important factor in drivers' route choice compared to other factors like travel time, free flow, road type considered in this thesis.

7.2 Scenarios applied to airport-city center routes

By now, a general route choice pattern of taxi drivers has been provided. However, it is still unclear whether this general pattern holds for specific cases. To answer this question, the following section make clusters of the study area and pick out the routes from the airport to the city center as a specific study. The following subsections describe in detail the clustering algorithm and how it is applied to select the area and the cases for this specific study.

7.2.1 DBSCAN

Because of heterogeneity of different travel ODs, it is possible that the above aggregate analysis may underestimate the effects of certain factors under some types of travels. In order to find representative cases confirming the above analysis, it is necessary to find places with high visiting frequencies and then extract routes from these places. One solution is to cluster the drop-off and pick-up locations of taxi drivers.

In order to find the representative places, Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was adopted here to identify the clusters of points (Ester et al., 1996). The DBSCAN algorithm takes in two parameters: the minimum number of points (minpts), and the range of clustering (ϵ). Because the clustering results largely depends on the settings of the two parameters, various combinations of the two parameter were explored to find a good set of parameter values.

Table 9 shows the number of cluster for different combinations of parameters. As the darkness of the colors in the cells indicates, the number of cluster decreases with the increasing of both minpts and ϵ. When the values of minpts and ϵ are low, for example minpts $< 400\%$ and $\epsilon < 300$, the numbers of clusters are high. Such high number of clusters is less helpful for identifying taxi pick-up patterns than a mild number of clusters. On the contrary, when both of the values are low, the algorithm is only able to identify a very small number of factors that makes various features of clusters indistinguishable.

						ϵ					
	N	100	200	300	400	500	600	700	800	900	1000
	100%	4227	1822	558	147	53	36	27	18	12	12
	200%	3204	1291	400	99	44	36	27	19	15	12
	300%	2585	929	296	77	36	20	18	15	10	9
	400%	2107	735	236	59	27	18	12	10	10	8
	500%	1746	594	206	70	24	11	10	8	7	6
	600%	1455	487	180	68	24	15	10	7	6	5
	700%	1227	409	154	66	28	15	7	9	6	5
m	800%	1050	345	145	66	22	18	11	7	5	4
i	900%	889	302	125	61	28	13	7	5	3	4
n	1000%	774	267	103	59	26	15	10	6	3	3
p	1100%	693	228	93	51	29	16	10	8	4	3
t	1200%	606	203	86	51	24	15	9	8	5	3
s	1300%	518	186	87	43	22	11	11	7	5	4
	1400%	453	158	77	41	29	13	11	9	7	6
	1500%	391	145	81	34	29	12	11	10	5	7
	1600%	352	130	79	35	25	11	11	9	7	5
	1700%	322	124	76	38	19	15	9	5	4	5
	1800%	282	104	68	29	19	15	8	8	4	2
	1900%	259	88	61	26	18	15	7	7	5	1
	2000%	229	83	58	30	16	14	8	5	4	2

Table 9. Number of cluster with different combinations of parameters (minpts in the first column and ϵ in the first row)

If too many clustered are included, it is hard to find the general patterns. In contrast, if only a few clusters are chosen, it is likely to omit a lot of information. Finally, by balancing between identifiability of patterns and availability of information, ϵ value of 500 and the minpts value of 600% (3000) were chosen to produce the final clusters. As shown in Figure 31, there are in total 24 places that were clustered. The largest cluster, as indexed with number 1, corresponds to the city center with busiest commercial, travel and financial activities. Cluster 2 corresponds to Lujiazui which is the financial center of Shanghai. Cluster 3, 6, 8, 17, 19, 20 and 24 are major residential areas. Cluster 4 surround the Shanghai Children Median Center where there are high demand for taxi services due to children's health emergencies. Cluster 5, 10, 12 and 22 center around universities and high schools. Cluster 7 and 9 are exhibition and park areas. Cluster 11 is the location of the city zoo. Cluster 13, 15 and 18 are shopping malls or shopping streets. Cluster 14 and 23 are airports. Cluster 16 is near the location of the city theater, and cluster 21 is the bank zone. Overall, clustered places represent a vast of purposes ranging from education to entertainment, from shopping to working in financial business. While most of the clusters are located within the city center, the only exception is the cluster 23, the location of the Pudong International Airport.

Therefore, the travel routes from the Pudong International Airport provides the opportunity to test whether the aggregate analysis is commonly hold for all types of travels with different ODs.

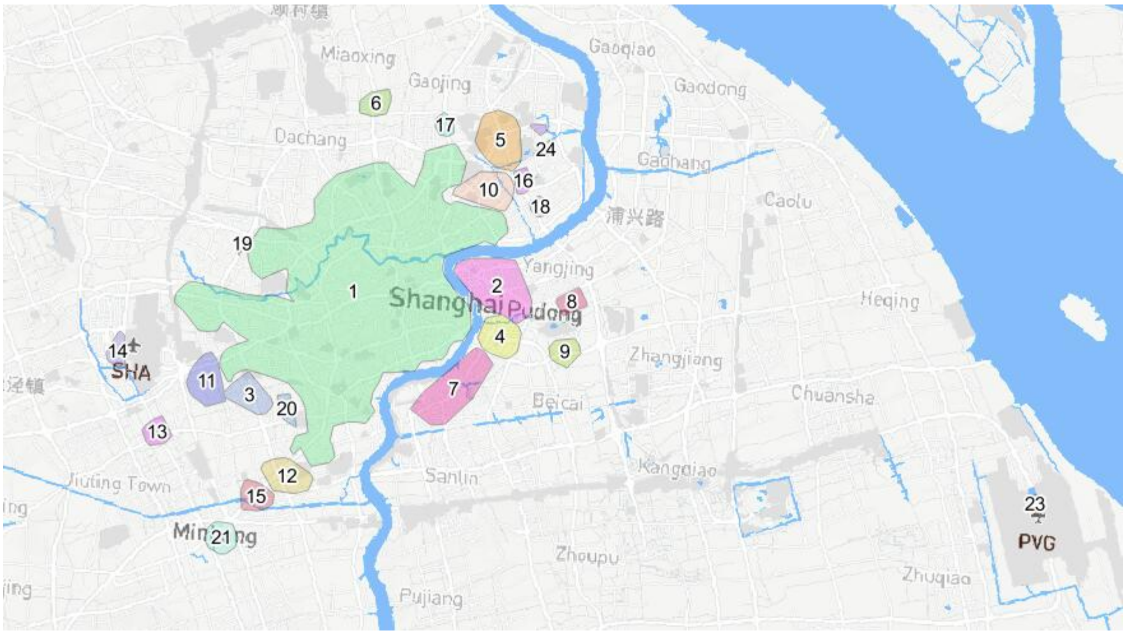


Figure 31. Clusters of most frequently visited places

7.2.2 Results

To examine the results of scenarios in applying to specific travel routes, the routes between the city center and the Pudong International Airport were chosen as a typical example.

By restricting the origins from the airport, it is possible to identify the clusters within the city center. And applying the DBSCAN algorithm again to the restricted dataset, the analysis results in the following clusters as shown in Figure 32. Cluster A is around the Nanjing Road, the most famous shopping street in Shanghai. Cluster B is the location of Lujiazui financial district. Cluster C is the hotel area. And cluster D points to Xujiahui, the commerce and cultural center of Shanghai.

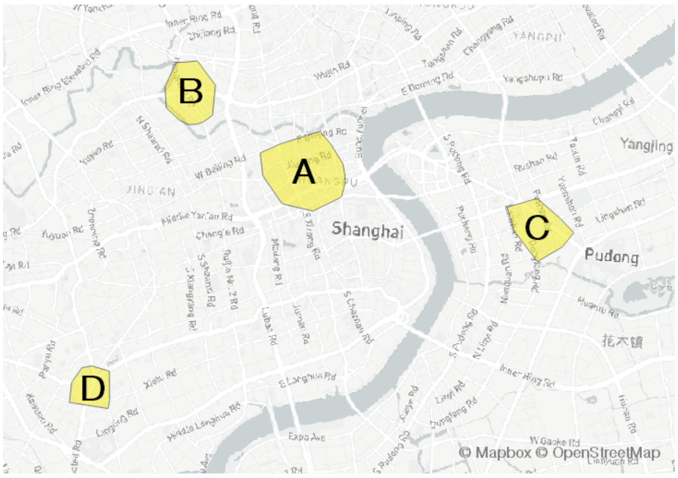


Figure 32. Clusters of most frequently visited places by routes from the Pudong International Airport

By extracting routes from the cluster of airport to the cluster of the center center, Figure 33 is able to visualize real routes from the airport to the city center. The transparency and darkness the the route color indicate the usage frequency. The darker the color, the more frequently are the routes used by taxi drivers. While several routes are used more frequently by taxi drivers traveling from the airport to the city center, there are some detours that travel around the city.

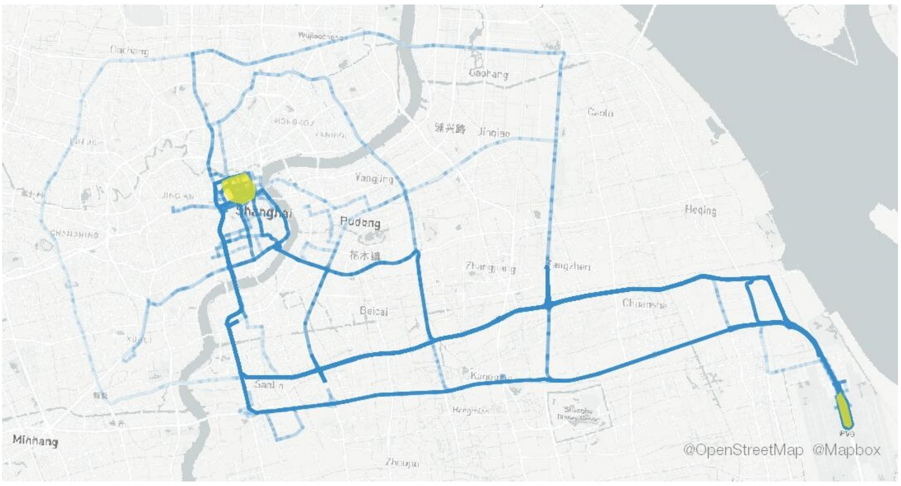


Figure 33. Routes from the Pudong International Airport to city center

By applying all the scenarios to the case of airport traffic, the following results were achieved and summarized by Table 10. The overall accuracy dropped dramatically in airport traffics, indicating more uncertainty of taxi drivers route choices when the travel distances increase. This is because with longer travel distance, taxi drivers face more choices, and therefore the proportion of overlap between the real and the shortest routes accordingly decreases.

Scenario	Weighted Criteria	Mean (%)	Over 75% match	Improvement
0	Minimal distance	13.25	0.00	
1	Least freeflow travel time	24.22	4.67	YES
2	Least observed travel time	48.00	12.62	YES
3	Minimal (road category * distance)	51.79	17.76	YES
4	Minimal (road category * time)	30.78	9.35	YES
5	Minimal (road usage frequency * distance)	32.34	0.00	YES
6	Minimal (road usage frequency * time)	25.84	5.14	YES
7	Minimal (turns * distance)	10.68	0.00	NO
8	Minimal (turns * time)	20.52	4.67	YES
9	Minimal (left turns * distance)	13.26	0.00	NO
10	Minimal (left turns * time)	24.20	4.66	YES

Table 10. Prediction accuracy of different scenarios to the routes from Pudong International Airport to the city center

For the trips from the airport to the city center, the freeflow travel time and the observed travel time play a more important role than in general trips as illustrated in the previous section. In particular, the prediction accuracy increases dramatically from 13% to 48% when taking into account of observed travel time. Road type is also important as shown in Scenario 3 and 4. Considering road types, the improvement of average prediction accuracy and the high-level matches (over 75% match) are significantly high. Road usage frequency also influence taxi drivers' choices but are less important in such case than travel time. Number of turns and number of left turns become even less significant as shown in Scenario 7 and 9. Even though in combined with travel time, they help to improve the prediction accuracy as in Scenario 8 and 10.

Overall, the improvement of prediction accuracy is mainly due to travel distance, travel time, road type and usage frequencies as quite similar to the general analysis in the previous section. Compared to the general analysis, the travel time between the airport and the city center plays a more important role.

Concluding from the above general and specific preference analyses, travel distance and travel time have the highest influence on taxi drivers' route choices, and the their relative impacts depend on how far the real routes are. When the real routes are long as in the case from the airport to the city center, travel time outplays travel distance and becomes the most important factor in drivers' choice. The impact of road type also increases with the increasing travel distances. Other factors including road usage frequency, number of turns and number of left turns have some impact on taxi drivers' route choice.

8 Conclusion, limitations and outlook

Tradition wisdom explaining drivers' route choice behavior focuses on simple factors such as the shortest routes or the fastest routes. The intuition behind such explanation is that drivers may enjoy high utilities when the costs of the travel, either in terms of time or distance, are low. However, the real chosen routes always deviate from the prediction using these simple factors. The reasons for the deviation from the prediction are multi-faceted. And one of the reasons is lacking sufficient data and efficient analytical tools to explore the relations between multiple factors and the final chosen routes.

Do taxi drivers choose the shortest routes? The answer from the above analyses using large-scale FCD is a partial YES. However, there are also other elements that a taxi driver concerns about when deciding the routes between the OD. From preference analysis, road type, travel speed limit, observed travel speed, frequency of road usage have impacts on the taxi drivers' route choice.

Map matching results show that taxi drivers are most active near the city center where there is a huge demand for taxi services from residents, tourists, office commuters, and public officers. There is also a high demand for taxi services in the airport where customers always expect punctuate travels and try to avoid delays. The concentration of taxi service activities in city center and around the airport provides valuable opportunity to study both the general and specific patterns of taxi drivers' route choice behavior.

Comparison analysis identifies several relations on different features between real routes and the shortest routes. In particular, calculation of PSL show that even though there is a large proportion of overlaps between real routes and the shortest routes (on average 46%), the actual deviations from the two are high as well. Real routes are on average longer than the shortest routes by 16% and are more circuitous. The difference between the real and the shortest routes difference between different travel distances: the longer the travel distance, the larger the difference between the two routes. Segment-based analysis also shows that real routes and the shortest routes differ in terms of road type, intersections and number of left turns.

Going a step further from comparison analysis, preference analysis identifies a set of scenarios to explore the extent to which different factors impact the final route choices of taxi drivers. Compared to the shortest-distance scenario with an average of 46.61% prediction accuracy, the scenario in combination with travel distance and road usage frequency improves the prediction accuracy to 53.56%. While, in the general situation travel distance still plays the major role, with the increase of travel distances, travel time becomes more important. As the trips from the airport to city center shows, observed travel time improves the prediction accuracy of real routes from 13.25% in shortest-distance scenario to 48%. Furthermore, the results also indicate that taxi drivers prefer to choose the most frequently used routes by other taxi drivers. In addition, compared to the original shortest distance prediction, the addition of road type to the shortest path algorithm also significantly improve the prediction results.

There are at least three limitations in this thesis. First, due to computational constraints, only the top 10 routes to predict the best matches in Scenario 7,8, 9 and 10 could be calculated for each real routes. Therefore, further work needs to be done to optimize the efficiency of the k-shortest-path algorithm so that it is computable with large-scale road networks. Second, the period of study include two weeks. However, it is possible to adapt the study period and include also other external events such as ceremonies and typhoon. Future research may extend the period of study and examine the impact of such events on the taxi drivers' behavior. Finally, this thesis doesn't consider the impact of context, such as weather and external events, on taxi drivers' choice. Future study may also extend the analysis to examine the taxi drivers' context awareness.

9 Bibliography

Agrawal, S., H. Zheng, S. Peeta, and A. Kumar (2016). Routing Aspects of Electric Vehicle Drivers and Their Effects on Network Performance. *Transportation Research Part D: Transport and Environment 46*, 246–266.

Arnott, R., A. de Palma, and R. Lindsey (1992). Route Choice with Heterogeneous Drivers and Group-specific Congestion Costs. *Regional Science and Urban Economics 22*(1), 71–102.

Bentley, J. L. and H. A. Maurer (1980, Feb). Efficient worst-case data structures for range searching. *Acta Informatica 13*(2), 155–168.

Bernstein, D. and A. Kornhauser (1998). An introduction to map matching for personal navigation assistants. *Working Paper*.

Ciscal-Terry, W., M. Dell'Amico, N. S. Hadjidimitriou, and M. Iori (2016). An Analysis of Drivers Route Choice Behaviour Using GPS Data and Optimal Alternatives. *Journal of Transport Geography 51*, 119–129.

de Carvalho, E. S., S. Rasouli, and H. Timmermans (2016). Use of Concepts of Regret and Disappointment to Model Dynamic Route Choice Behavior under Uncertainty. *Transportation Research Board 95th Annual Meeting 2567*, 131–138.

De Fabritiis, C., R. Ragona, and G. Valenti (2008). Traffic Estimation and Prediction Based on Real Time Floating Car Data. *Proceedings of the 11th International IEEE Annual Conference on Intelligent Transportation Systems (ITSC 2013)*, 197–203.

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

El Najjar, M. E. and P. Bonnifait (2005). A road-matching method for precise vehicle localization using belief theory and kalman filtering. *Autonomous Robots 19*(2), 173–191.

Eppstein, D. (1998). Finding the k Shortest Paths. *SIAM Journal of Computing 28*(2), 652–673.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. pp. 226–231. AAAI Press.

Floyd, R. W. (1962, June). Algorithm 97: Shortest path. *Commun. ACM 5*(6), 345–.

Gan, H.-c., X. Ye, and Q. Wang (2010). Investigating the Effect of Travel Time Variability on Drivers' Route Choice Decisions in Shanghai, China. *Transportation Planning and Technology 33*(8), 657–669.

Gelenbe, E. and G. Sakellari (Eds.) (2012). *Computer and Information Sciences II*. London: Springer.

Golledge, R. G. (1995). Path Selection and Route Preference in Human Navigation: A Progress Report. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 988*, 207–222.

Greenfeld, J. S. (2002). Matching gps observations to locations on a digital map. In *81th Annual Meeting of the Transportation Research Board*, Volume 1, pp. 164–173.

Gühnemann, A., R.-P. Schäfer, K.-U. Thiessenhusen, and P. Wagner (2004). Monitoring Traffic and Emissions by Floating Car Data. *Working Paper*, 1–17.

Guo, R. Y. and H. J. Huang (2016). A Discrete Dynamical System of Formulating Traffic Assignment: Revisiting Smith's model. *Transportation Research Part C: Emerging Technologies 71*, 122–142.

He, X. and S. Peeta (2016). A Marginal Utility Day-to-day Traffic Evolution Model Based on One-step Strategic Thinking. *Transportation Research Part B: Methodological 84*, 237–255.

Hoffman, W. and R. Pavley (1959). A method for the solution of the n th best path problem. *Journal of the ACM (JACM) 6*(4), 506–514.

Hofmann-Wellenhof, B., H. Lichtenegger, and J. Collins (2012). *Global Positioning System: Theory and Practice*. Springer Science & Business Media.

Huber, W., M. Lädke, and R. Ogger (1999). Extended floating-car data for the acquisition of traffic information. In *Proceedings of the 6th World Congress on Intelligent Transport Systems*, pp. 1–9.

Ishii, H. (1978). A New method Finding the K-th Best Path in A Graph. *Journal of the Operations Research 21*(4), 469–475.

Jiang, X., Y. Ji, M. Du, and W. Deng (2014). A Study of Driver's Route Choice Behavior Based on Evolutionary Game Theory. *Computational Intelligence and Neuroscience 2014*, 1–10.

Jonker, R. and A. Volgenant (1987). A Shortest Augumenting Path Algorithm for Dense and Sparse Linear Assignment Problems. *Computing 340*, 325–340.

Kerner, B., C. Demir, R. Herrtwich, S. Klenov, H. Rehborn, M. Aleksic, and A. Haug (2005). Traffic state detection with floating car data in road networks. In *Intelligent Transportation Systems, 2005. Proceedings. 2005 IEEE*, pp. 44–49. IEEE.

Kim, S. K. (1998). Development of a map matching algorithm for car navigation system using fuzzy q-factor algorithm. In *Towards the new horizon together. Proceedings of the 5th world congress on intelligent transport systems, held 12-16 October 1998, Seoul, Korea. Paper no. 4020.*

Krakiwsky, E. J., C. B. Harris, and R. V. Wong (1988). A kalman filter for integrating dead reckoning, map matching and gps positioning. In *Position Location and Navigation Symposium, 1988. Record. Navigation into the 21st Century. IEEE PLANS'88., IEEE*, pp. 39–46. IEEE.

Li, Q., T. Zhang, H. Wang, and Z. Zeng (2011). Dynamic Accessibility Mapping using Floating Car Data: A Network-constrained Density Estimation Approach. *Journal of Transport Geography 19*(3), 379–393.

Manley, E. J., J. D. Addison, and T. Cheng (2015). Shortest Path or Anchor-based Route Choice: A Large-scale Empirical Analysis of Minicab Routing in London. *Journal of Transport Geography 43*, 123–139.

Miller-Hooks, E. D. and H. S. Mahmassani (2000). Least Expected Time Paths in Stochastic, Time-varying Transportation Networks. *Transportation Science 34*(2), 198–215.

Newson, P. and J. Krumm (2009). Hidden Markov Map Matching through Noise and Sparseness. *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09*, 336.

Ochieng, W. Y., M. Quddus, and R. B. Noland (2003). Map-matching in complex urban road networks. *Revista Brasileira de Cartografia 2*(55).

Palma, A. D. E. and D. Rochat (1999). Understanding Individual Travel Decisions: Results from A Commuters Survey in Geneva. *Transportation 26*, 263–281.

Papinski, D. and D. M. Scott (2011). A GIS-based Toolkit for Route Choice Analysis. *Journal of Transport Geography 19*(3), 434–442.

Papinski, D. and D. M. Scott (2013). Route Choice Efficiency: An Investigation of Home-to-work Trips Using GPS Data. *Environment and Planning A 45*(2), 263–275.

Paulin, T. and S. Bessler (2013). Controlled Probing - A System for Targeted Floating Car Data Collection. *Proceedings of the 16th International IEEE Annual Conference on Intelligent Transportation Systems (ITSC 2013)*, 1095–1100.

Phuyal, B. P. (2002). Method and use of aggregated dead reckoning sensor and gps data for map matching. In *ION GPS 2002: 15 th International Technical Meeting of the Satellite Division of The Institute of Navigation.*

Prato, C. G. and T. K. Rasmussen (2016). Estimating the Value of Congestion under Road Pricing Schemes. *Australasian Transport Research Forum 2016 Proceedings* (November), 1–13.

Prato, C. G., T. K. Rasmussen, and O. A. Nielsen (2014). Estimating Value of Congestion and of Reliability from Observation of Route Choice Behavior of Car Drivers. *Transportation Research Record: Journal of the Transportation Research Board* (January 2016), 20–27.

Pyo, J.-S., D.-H. Shin, and T.-K. Sung (2001). Development of a map matching method using the multiple hypothesis technique. In *Intelligent Transportation Systems, 2001. Proceedings. 2001 IEEE*, pp. 23–27. IEEE.

Quddus, M. A., W. Y. Ochieng, and R. B. Noland (2007). Current Map-matching Algorithms for Transport Applications: State-of-the Art and Future Research Directions. *Transportation Research Part C 15*, 312–328.

Quddus, M. A., W. Y. Ochieng, L. Zhao, and R. B. Noland (2003). A general map matching algorithm for transport telematics applications. *GPS solutions 7*(3), 157–167.

Ramaekers, K., S. Reumers, G. Wets, and M. Cools (2013). Modelling Route Choice Decisions of Car Travellers Using Combined GPS and Diary Data. *Networks and Spatial Economics 13*, 351–372.

Razo, M. and S. Gao (2013). A Rank-dependent Expected Utility Model for Strategic Route Choice with Stated Preference Data. *Transportation Research Part C: Emerging Technologies 27*, 117–130.

Schafer, R.-P., K.-U. Thiessenhusen, and P. Wagner (2002). A Traffic Information System by Means of Real-Time Floating-Car Data. *Working Paper*, 1–8.

Shang, W., K. Han, W. Ochieng, and P. Angeloudis (2016). Agent-Based Day-to-Day Traffic Network Model with Information Percolation. *Transportmetrica A: Transport Science*.

Simroth, A. and H. Zähle (2011). Travel Time Prediction Using Floating Car Data Applied to Logistics Planning. *IEEE Transactions on Intelligent Transportation Systems 12*(1), 243–253.

Smith, M. J. (1979). The Existence, Uniqueness and Stability of Traffic Equilibria. *Transportation Research Part B 13*(4), 295–304.

Smith, M. J. (1983). The Existence and Calculation of Traffic Equilibria. *Transportation Research Part B 17*(4), 291–303.

Smith, M. J. (1984). The Stability of A Dynamic Model of Traffic Assignment – An Application of A Method of Lyapunov. *Transportation science 18*(3), 245–252.

Sun, D. J., C. Zhang, L. Zhang, F. Chen, and Z.-R. Peng (2014). Urban Travel Behavior Analyses and Route Prediction Based on Floating Car Data. *Transportation Letters 6*(3), 118–125.

Tadic, B., G. J. Rodgers, and S. Thurner (2007). Transport on Complex Networks: Flow, Jamming and Optimization. *International Journal of Bifurcation and Chaos 17*(7), 2363–2385.

Tang, J., H. Jiang, Z. Li, M. Li, F. Liu, and Y. Wang (2016). A Two-layer Model for Taxi Customer Searching Behaviors Using GPS Trajectory Data. *IEEE Transactions on Intelligent Transportation Systems 17*(11), 3318–3324.

Von Neumann, J. and O. Morgenstern (1944). *Theory of Games and Economic Behavior.* Princeton: Princeton University Press.

Wardrop, J. G. (1952). Some Theoretical Aspects of Road Traffic Research. *Proceedings of the Institution of Civil Engineers 36*, 325–378.

White, C. E., D. Bernstein, and A. L. Kornhauser (2000). Some Map Matching Algorithms for Personal Navigation Assistants. *Transportation Research Part C 8*, 91–108.

Wilson, R. J. (1996). *Introduction to Graph Theory (Fourth Edition).* Essex: Longman.

Yan, G., T. Zhou, B. Hu, Z.-q. Fu, and B.-h. Wang (2006). Efficient Routing on Complex Networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 1–5.

Yang, F. and D. Zhang (2009). Day-to-day Stationary Link Flow Pattern. *Transportation Research Part B 43*(1), 119–126.

Yang, Y., E. Yao, Z. Yang, and R. Zhang (2015). Modeling the Charging and Route Choice Behavior of BEV Drivers. *Transportation Research Part C: Emerging Technologies 65*, 190–204.

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

Yen, J. Y. (1971). Finding the K Shortest Loopless Paths in A Network. *Management Science 17*(11), 712–716.

Yu, M. (2006). *Improved Positioning of Land Vehicle in ITS Using Digital Map and Other Accessory Information.* Ph. D. thesis, The Hong Kong Polytechnic University.

Zhao, Y. (2010). *Vehicle Location and Navigation Systems.* Boston [u.a.]: Artech House.

Ziliaskopoulos, A. K. and H. S. Mahmassani (1993). Time-Dependent, Shortest-Path Algorithm for Real-Time Intelligent Vehicle Highway System Applications. *Transportation Research Record* (2), 94–100.