

# STARS

University of Central Florida  
**STARS**

---


Electronic Theses and Dissertations, 2020-

---

2020

## Traffic Speed Prediction and Mobility Behavior Analysis Using On-Demand Ride-Hailing Service Data

Jiechao Zhang  
*University of Central Florida*

 Part of the [Civil Engineering Commons](#), and the [Transportation Engineering Commons](#)  
Find similar works at: <https://stars.library.ucf.edu/etd2020>  
University of Central Florida Libraries <http://library.ucf.edu>

This Masters Thesis (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2020- by an authorized administrator of STARS. For more information, please contact [STARS@ucf.edu](mailto:STARS@ucf.edu).

---

### STARS Citation

Zhang, Jiechao, "Traffic Speed Prediction and Mobility Behavior Analysis Using On-Demand Ride-Hailing Service Data" (2020). *Electronic Theses and Dissertations, 2020-*. 158.  
<https://stars.library.ucf.edu/etd2020/158>



**TRAFFIC SPEED PREDICTION AND MOBILITY BEHAVIOR  
ANALYSIS USING ON-DEMAND RIDE-HAILING SERVICE DATA**

by

**JIECHAO ZHANG**  
B.Sc. Beijing Jiaotong University, 2015

A thesis submitted in partial fulfillment of the requirements  
for the degree of Master of Science  
in the Department of Civil, Environmental and Construction Engineering  
in the College of Engineering and Computer Science  
at the University of Central Florida  
Orlando, Florida

Spring Term  
2020

Major Professor: Samiul Hasan

© 2020 Jiechao Zhang

## ABSTRACT

Providing accurate traffic speed prediction is essential for the success of Intelligent Transportation Systems (ITS) deployments. Accurate traffic speed prediction allows traffic managers take proper countermeasures when emergent changes happen in the transportation network. In this thesis, we present a computationally less expensive machine learning approach XGBoost to predict the future travel speed of a selected sub-network in Beijing's transportation network. We perform different experiments for predicting speed in the network from future 1 min to 20 min. We compare the XGBoost approach against other well-known machine learning and statistical models such as linear regression and decision tree, gradient boosting tree, and random forest regression models. Three metrics MAE, MAPE, and RMSE are used to evaluate the performance of the selected models. Our results show that XGBoost outperforms other models across different experiment conditions. Based on the prediction accuracy of different links, we find that the number of vehicles operating in a network also affect prediction performance.

In addition, understanding individual mobility behavior is critical for modeling urban dynamics. It provides deeper insights on the generative mechanisms of human movements. Recently, different types of emerging data sources such as mobile phone call detail records, social media posts, GPS observations, and smart card transactions have been used to analyze individual mobility behavior. In this thesis, we report the spatio-temporal mobility behaviors using large-scale data collected from a ride-hailing service platform. Based on passenger-level travel data, we develop an algorithm to identify users' visited places and the functions of those places. To characterize temporal movement patterns, we reveal the differences in trip generation characteristics between commuting and non-commuting trips and the distribution of gap time between consecutive trips. To understand spatial mobility patterns, we observe the distribution of the number of visited place

and their rank, the spatial distribution of residences and workplaces, and the distribution of travel distance and travel time. Our analysis highlights the differences in mobility patterns of the users of ride-hailing services, compared to the findings of existing mobility studies based on other data sources. Our study shows a tremendous potential of developing high-resolution individual-level mobility model that can predict the demand of emerging mobility services with high accuracy.

Keywords: Traffic prediction, Mobility Behavior, Machine learning, Ride-hailing services.

## **ACKNOWLEDGMENTS**

I would like to convey my heartiest gratitude to my honorable supervisor Dr. Samiul Hasan for his excellent supervision and constant support in this thesis. I would like to acknowledge the support and encouragement from my family and friends.

## TABLE OF CONTENTS

LIST OF FIGURES .....	viii
LIST OF TABLES .....	ix
CHAPTER 1: INTRODUCTION .....	1
1.1 Introduction .....	1
1.2 Thesis Contribution .....	3
1.3 Objectives of the Thesis .....	4
1.4 Thesis Organization.....	5
CHAPTER 2: LITERATURE REVIEW .....	6
2.1 Traffic Prediction .....	6
2.2 Mobility Behavior Analysis .....	8
CHAPTER 3: SHORT-TERM TRAFFIC SPEED PREDICTION .....	11
3.1 Introduction and motivation .....	11
3.2 Dataset and Modeling Framework .....	12
3.2.1 Dataset Description.....	12
3.2.2 Modeling Framework.....	14
3.3 Results .....	17
3.4 Discussion .....	23
CHAPTER 4: MOBILITY BEHAVIOR ANALYSIS .....	25
4.1 Introduction .....	25
4.2 Data and Methods.....	26
4.2.1 Dataset Description.....	26
4.2.2 Distance-Based Visited Place Generation Algorithm.....	27
4.3 Empirical Results .....	30
4.3.1 Temporal Pattern - Trip Generation.....	30
4.3.2 Gap Time .....	32
4.3.3 Number of Visited Places and their Rank.....	33
4.3.4 Distribution of Travel Distance .....	35
4.3.5 Distribution of Travel Distance .....	37
4.4 Discussion .....	40
4.5 Implications .....	43
CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS .....	45
5.1 Summary .....	45

5.2	Limitations and Future Research Direction .....	46
REFERENCES	.....	48



## LIST OF FIGURES

Figure 3.1: Study Area and Map Matching Algorithm Examples .....	14
Figure 3.2: Traffic performance in different time periods .....	17
Figure 3.3: Performance of Different Prediction Methods .....	20
Figure 3.4: Ground Truth vs. Predictions for different horizons using XGBoost .....	20
Figure 3.5: Performance of Prediction Methods of Different Links.....	22
Figure 3.6: Prediction performance vs. the number of records.....	22
<b>Figure 4.1:</b> The probability of the demand of ride-hailing service (March 1-June 30, 2017): (a) the distribution of the number of daily trips using the ride-hailing service ('M' means Monday of every week) (b) the distribution of average hourly trips number of ride-hailing service. ....	32
<b>Figure 4.2:</b> The distribution of gap time. (a) the average gap time (min) vs. their corresponding users' groups; (b) the probability of different gap time (min) in two scales .....	33
<b>Figure 4.3:</b> The distribution of passengers' visited place number and the probability of visited place rank: (a) the probability of different visited place number. (b) the probability of visiting different places based on its rank level in log-log scale .....	35
<b>Figure 4.4:</b> The distributions of travel distance per trip with fitting curves of selected distributions: (a) distribution of travel distance (km) per trip for commuting trips; (b) distribution of travel distance (km) per trip for non-commuting trips.....	37
<b>Figure 4.5:</b> The distribution of travel time per trip with fitting curves of selected distributions: (a) Distribution of travel time per trip for commuting trips. (b) distribution of travel time per trip for non-commuting trips.....	39

## LIST OF TABLES

Table 3.1: An Example of the Trajectory Data.....	13
<b>Table 3.2:</b> Values of Parameters for Different Machine Learning Methods .....	19
Table 3.3: Training Time of Different Models .....	21
Table 4.1: Detailed Data Attributes .....	27
Table 4.2: The results of K-S test of selected distribution for travel distance.....	37
Table 4.3: The results of K-S test of selected distribution for travel time.....	39

# CHAPTER 1: INTRODUCTION

## 1.1 Introduction

Short-term prediction of traffic states is core driving force of implementation of intelligent transportation system (Ma et al., 2015). With the increasingly worsen congestion problem in cities, accurately forecasting the traffic states will provide information to the individuals so that they can set up efficient travel routes by avoiding the traffic jams. Besides, the prediction of traffic states can also bring innovative insights to the traffic managers, policy makers and transportation planners, which can be beneficial to come up with advanced countermeasures to mitigate the congestion problems (Yao et al., 2017).

In addition, the patterns of individual mobility behavior have significant influence on the operation of transportation network. Human movement is one of the critic aspects in the dynamic of urban mobility. Understanding the spatio-temporal individual mobility behavior will provide new ideas to solve many transportation problems, such as urban planning (Sun et al., 2016) and traffic management (Chen et al., 2016). In addition, revealing the individual mobility behavior patterns shows great potential in developing high-resolution generative models which can be used to establish agent-based trip activity simulation.

However, both the short-term forecasting of traffic states and the analysis of individual mobility behavior are challenging in the past due to a lack of massive data. To solve these two problems, we have to consider both the spatial and temporal dependencies, which may affect the operation of transportation, because the individuals have high regional and temporal periodicity (Ke et al., 2017). In terms of the traffic states prediction, the previous researches always utilize the data extracted from traffic sensors, which are lower efficiency and higher cost (Ma et

al., 2015). To uncover the individual mobility behavior patterns, the existing studies focus on different types of data sources, such as bank notes (Brockmann et al., 2006), social media data (Jurdak et al., 2015), mobile phone data (Huang et al., 2018) and smart card data (Zhao et al., 2018). Nowadays, with the rapid development of the innovative emerging transportation service – on-demand ride-hailing service, a more advanced type of data is available for transportation analytics and predictions.

In recent years, the emergence on-demand ride-hailing service, such as Uber, Lyft and DiDi, has significantly affected the human movement behaviors with the rapid development of mobile internet (Ke et al., 2017). The on-demand ride-hailing service enables individuals to call for an urban mobility service via smart phone applications everywhere in a city (Chen et al., 2017), which can make up the shortcomings of traditional taxicab service with higher empty-loaded rate. With massive users of on-demand ride-hailing service platforms, we can efficiently obtain a large amount of transportation data which can be used for traffic states prediction and revealing the individual mobility behavior, which is unbelievable before.

In this thesis, we present a new state-of-the-art machine learning model XGBoost to predict the future traffic speed of a road network using high-resolution trajectory data extracted from on-demand ride-hailing service vehicles and report the findings on mobility behavior of the new population group—users of emerging on-demand ride-hailing services—after analyzing large-scale trip data from a ride-hailing platform.

## 1.2 Thesis Contribution

This study has made several contributions towards traffic operations and management by developing short-term traffic prediction models and analyzing individual mobility behavior of on-demand ride-hailing service users.

In terms of the short-term traffic speed prediction, we present a new state-of-the-art machine learning model (XGBoost) to predict the traffic speed of a network using high-resolution trajectory data extracted from Didi—an on-demand ride-hailing service platform. The main contributions of this section are as follows:

- We use high-resolution trajectory data, collected from emerging ride-hailing services, to predict the future traffic states at a network level.
- We adopt a machine learning approach that is computationally less expensive, and hence suitable for real-time applications. As such, it offers new results by comparing other machine learning approaches commonly used for prediction problems.
- We utilize new variables such as spatial (link information) and service vehicle flow (number of vehicles operating in the network) to predict future network states. Analyzing the relationship between prediction accuracy and service vehicle as a predictor variable, it offers new insights for similar prediction problems.

For the mobility behavior analysis, to the best of our knowledge, this is the first study that reveals individual mobility patterns of ride-hailing service users based on large-scale data available from a ride-hailing platform. The main contributions of this section are as follows:

- We reveal the spatio-temporal mobility patterns of ride-hailing service users. Although ride-hailing platforms have been serving demand for several years, previous studies did not investigate the mobility patterns of the users of these services.
- We investigate critical aspects of on-demand ride hailing services such as the gap time between two consecutive rides and the rank of visited places of the users.
- We fit the distribution of travel distance and travel time of on-demand ride-hailing service trips and the results can be used to establish activity generation mechanisms for agent-based demand simulations which will significantly benefit the operations and management of on-demand ride-hailing services as well as urban planning and traffic management.

### **1.3 Objectives of the Thesis**

The focus of this study is to utilize the data extracted from on-demand ride-hailing service platform to predict the traffic network status and analyze the individual mobility behavior. Two different types of dataset – trajectory data and origin-destination (OD) data – have been used in this thesis.

The main objectives of this study are:

- To develop a machine learning model to predict the traffic state (speed) considering spatial and temporal dependency of the traffic pattern using high-resolution floating car trajectory data.
- To evaluate the performance of prediction model in traffic prediction and compare it with traditional machine learning and statistical models.
- To reveal the spatio-temporal patterns of individual mobility behavior of on-demand ride-hailing service users.

## **1.4 Thesis Organization**

The rest of the thesis is organized as follows: Chapter 2 provides literature review on the traffic prediction and mobility behavior analysis. Chapter 3 provides the data description, analysis, methodology, and result for short term traffic speed prediction. Chapter 4 describes the data description, methodology, and result for individual mobility behavior analysis. Chapter 5 presents the summary and conclusions of the thesis.

## **CHAPTER 2: LITERATURE REVIEW**

### **2.1 Traffic Prediction**

Predicting short-term traffic speed at different levels such as road segments or networks, plays an essential role in traffic management and route choice applications. However, prediction problems have been challenging to solve since achieving a high accuracy level is required for these applications (Zhao et al., 2019). Previously, various studies have been made to predict traffic speeds using statistical or machine learning methods such as support vector regression (SVR) (Wu et al., 2004), k-nearest neighbor regression (KNN) (Xia et al., 2016), time series models (seasonal ARIMA) (Williams and Hoel, 2003) and variants of neural network model such as artificial neural networks (ANN) (Huang and Ran, 2003), convolutional neural networks (Ma et al., 2017a), and long short term network models (Rahman and Hasan, 2018).

Previous studies developed multiple short-term traffic prediction models based on previous traffic patterns. For instance, in the study (Myung et al., 2011), the researchers developed a k-nearest neighbor regression model using the data extracted from traffic detectors. In addition, using multivariate non-parametric regression model, a researcher forecasted on the travel time for London railways with high performance. Baozhen Yao et al. (Yao et al., 2017) developed a support vector machine (SVM) model to predict the short-term traffic speed for an urban corridor considering both the spatial and temporal dependencies. They evaluated the performance of the model via empirical experiments on the GPS data of taxi and found that the SVM outperforms other machine learning models such as ANN and KNN. Studies also applied time series models to predict the short-term traffic speed. Wang et al. (Wang et al., 2014) developed a hybrid empirical mode decomposition and autoregressive integrated moving average (ARIMA) model to forecast



the freeways' speed, which has a good performance. Besides, Barimani et al. (Barimani et al., 2012) used Kalman Filter based model to predict the short-term traffic speed with the data extracted from the transportation detector which show better performance than the multilayer feed neural network.

Additionally, neural networks have been found to provide highly accurate traffic speed prediction. Hinsbergen (Van Hinsbergen et al., 2011) utilized Bayesian neural networks to forecast travel time in urban networks and proved that the neural networks can predict the so-called 'low-frequency trend'. From then on, many novel works based on neural networks have been proposed. For example, to capture the nonlinear patterns of transportation, Ma et al. (Ma et al., 2015) developed an long short-term neural network (LSTM) model for traffic speed prediction using the data extracted from traffic microwave detectors which shows higher accuracy than the other parametric and nonparametric models. Rahman et al. (Rahman and Hasan, 2018) adopted a Long Short-Term Memory Neural Network (LSTM-NN) to predict traffic speed in freeways under extreme traffic demand, such as a hurricane evacuation. In addition, to predict the short-term arterial speed, Jia et al. (Jia et al., 2016) established a deep belief network model, which is trained via a greedy unsupervised algorithm, showing better performance than the back propagation neural network and time series model (e.g. ARIMA). Besides LSTM, Ma et al. (Ma et al., 2017a) developed a convolutional neural network model to predict large-scale transportation network. In the model, the transportation network is seen as an image and the model outperforms other machine learning models. However, although neural network based models show high performance for prediction problems, lack of interpretability, longer training time, and required parameter tuning limit the application of these models (Zhang et al., 2016).

In recent years, ensemble algorithms have become highly useful for solving prediction and classification problems in many different fields. In particular, tree-based ensemble algorithms (Zhou, 2012) have become very successful, with XGBoost as one of the most powerful and efficient tree-based ensemble algorithms. In our study, we have applied the XGBoost model to predict the future traffic speed of links and evaluated the model performance based on floating car trajectory data. The XGBoost algorithm has outperformed other machine learning models indicating its high potential for traffic prediction problems.

## **2.2 Mobility Behavior Analysis**

Individual mobility behavior reflects the spatio-temporal dynamics of urban mobility at a high resolution (Brockmann et al., 2006). Traditionally, survey-based travel data have been used to analyze and model individual mobility behavior. Although travel survey methods have evolved from traditional pen-and-paper based data collection to nowadays web and smartphone-based data collection (Wolf et al., 2001) approaches, high cost and low sample size are major challenges towards implementing these tools at scale (Wu et al., 2011). To overcome the limitations travel survey data, researchers have started utilizing novel data sources such as bank notes (Brockmann et al., 2006) and mobile phone data (Gonzalez et al., 2008) to analyze human movement statistics for understanding individual mobility behavior.

With widespread adoption of mobile phones and location-based services, various large-scale high-resolution datasets with varying capabilities have been used to understand individual mobility behavior (Alessandretti et al., 2017, Zhang et al., 2018). For instance, call details records (CDR) from mobile phones can provide deep insights on individual mobility at a scale that was unimaginable before (Gonzalez et al., 2008, Chen et al., 2016, Huang et al., 2018). However, CDR

data are generated when a person makes a phone call or sends a message. It is a challenging task to predict when and where an individual will use his/her phone, which may result into incomplete travel information. Thus, inferring the origin and destination of individual activity is difficult based on such data (Huang et al., 2018). Social media data is another interesting data source that can provide rich information on individual travel and activity behavior (Rashidi et al., 2017). Through mining geo-location data recorded when user's check-in on social media platforms, individual activities can be identified over a long period, offering useful insights on individual travel patterns. However, these data do not include the precise start and end time of a trip, limiting applications in transportation (Hasan and Ukkusuri, 2018). Data from social media and mobile phones are defined as extrinsic mobility that do not directly observe individual travel behavior (Zhao et al., 2018).

Different from extrinsic mobility data (Zhao et al., 2018), smart card data (Hasan et al., 2013a) and floating car data (FCD) (Ehmke et al., 2012) can be defined as intrinsic mobility data that are directly collected from transportation system operations. For instance, smart card data are extracted from public transit operations, while FCD are collected from taxicab. Both types of datasets record when and where a user takes public transit (e.g., subway or bus) or taxi for a trip—giving precise information on the origin, destination, distance, price, and time of a trip. Unlike extrinsic data, intrinsic mobility data can offer mode-specific complete trajectory information, giving a different perspective to understand individual travel behavior.

Compared to smart card data, taxicab data have limitations to uncover individual mobility patterns, because passengers always pay in cash or credit cards when they take a taxi, without requiring the system to record individual details for historical tracking. Thus, most of the previous studies on human mobility under taxicab service focus on general urban resident's travel patterns

(Zheng et al., 2018) or the taxi drivers' travel behavior (Leng et al., 2016), instead of passenger mobility patterns. Due to the lack of available data, studies on individual mobility patterns using taxicab services hardly exist. However, a deeper understanding of individual mobility patterns under taxicab services from a passenger's perspective is significantly beneficial to many problems involving emerging ride-hailing services such as real-time demand prediction (Ke et al., 2017), designing ride-sharing operations (Alonso-Mora et al., 2017a), and designing mobility services for autonomous vehicles (Alonso-Mora et al., 2017b).

The emergence of on-demand ride-hailing platforms provides an innovative transportation service that can be easily requested via a smartphone app—providing longitudinal mobility data at an individual level (Contreras and Paz, 2018). These ride-hailing service platforms have a great potential in revealing individual mobility behaviors since the locations and timings of individual trips can be recorded through the GPS devices in the smartphone and stored in the platform. However, previous studies mainly used ride-hailing data for analyzing aggregate mobility behavior (Dong et al., 2018) and solving traffic modeling and prediction problems (Ke et al., 2017). As such, human mobility literature lacks an understanding of individual-level mobility patterns based on ride-hailing service data. To fill this research gap, we use large-scale data extracted from a major ride-hailing platform to analyze the mobility patterns of its users. To the best of our knowledge, this is the first research that presents individual mobility behavior from on-demand ride-hailing service data. The results of this research will provide valuable insights for many future studies such as demand prediction, policy making, and design of ride-hailing services.

## **CHAPTER 3: SHORT-TERM TRAFFIC SPEED PREDICTION**

### **3.1 Introduction and Motivation**

Predicting traffic states is one of the critical aspects of deploying intelligent transportation systems. Accurate and reliable prediction of travel states such as travel time enables people to make informed travel decisions. Real-time traffic predictions also help different navigation devices/applications recommend the best routes to travelers. In addition, real-time predictions of traffic states allow traffic managers to take pro-active countermeasures to avoid excessive traffic congestion or manage incidents.

However, predicting the states of large-scale transportation networks in real time is a challenging task since it needs massive information extracted from various sources or requires deployment of a large number of traffic sensors in the network (Ma et al., 2017a, Rahman and Hasan, 2018). However, with the adoption of GPS technologies and smartphone devices, we can observe vehicle trajectories in a road network in low cost with high accuracy—which enables efficient monitoring of traffic operations of large networks. In particular, thousands of taxis are driving on city roads every day (Wen et al., 2014). High-resolution trajectories collected from the GPS devices or navigation apps of these vehicles allow us to predict the spatio-temporal dynamics of traffic operations of an entire network, which was difficult in the past due to the lack of appropriate data (Dowling et al., 2004). Previous studies on taxicab data mostly focused on the origin-destination information for different transportation applications such as urban link travel time estimation (Zhan et al., 2013, Wu et al., 2004, Xia et al., 2016), spatial variation of urban taxi

ridership (Qian and Ukkusuri, 2015), and sparse trajectory information for traffic condition estimation (Herring et al., 2010).

In recent years—with widespread adoption of online ride-hailing services such as Uber, Lyft, and Didi—big data are being generated from these services in every second and high frequency trajectory data are available. But new methods are needed to utilize these data to estimate and predict traffic states of large-scale networks in real time. For this study, high-resolution trajectory data extracted from the Didi platform are used for speed prediction. The trajectory data of the Didi service vehicle are collected in every 12 seconds—providing high-resolution information to accurately forecast future traffic states.

Although previous studies have predicted traffic speeds using different approaches (Ma et al., 2017a, Rahman and Hasan, 2018, Wu et al., 2004, Xia et al., 2016), they hardly consider the training time of different models. When predicting the speed in a large network, training time can significantly influence the deployment of the models. Therefore, a fast and accurate model is required for traffic speed prediction in a road network level.

In this section, we present a new state-of-the-art machine learning model (XGBoost) to predict the traffic speed of a network using high-resolution trajectory data extracted from Didi—an on-demand ride-hailing service platform.

## **3.2 Dataset and Modeling Framework**

### **3.2.1 Dataset Description**

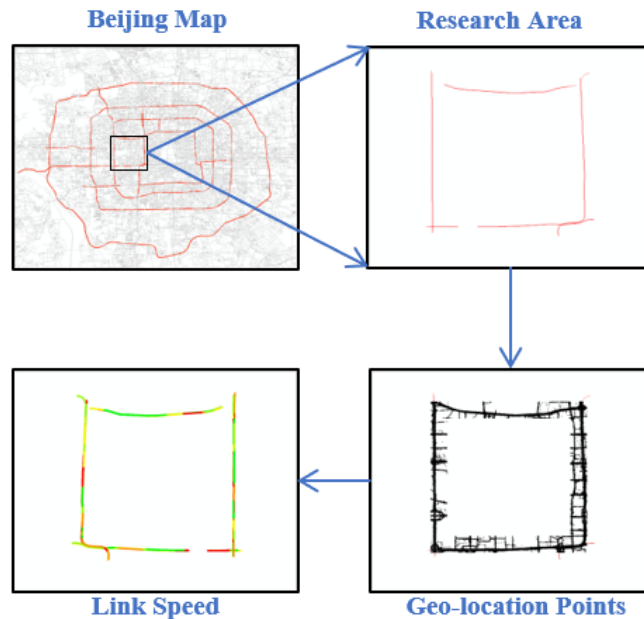
In this study, we have collected high-resolution vehicle trajectory data extracted from Didi, which is a ride-hailing service platform operating in Beijing, China. The data contains the basic information of coordinates, travel speed (km/h) and timestamp (MM/DD/YYYY HH: MM: SS)

(see Table 3.1). The time range of the dataset is 2:00 am to 1:00 pm on August 1, 2017. The original data were encoded by ASCII and stored in csv files. The total size of the data is more than 50 GB.

**Table 3.1:** An Example of the Trajectory Data

Record_ID	Time_Stamp	Longitude	Latitude	Speed
17594554834261	08/01/2017 2:07:05	116.3564	39.90442	10.8309
17594554839261	08/01/2017 2:07:08	116.3874	39.87759	16.6300
17594554840669	08/01/2017 2:07:06	116.3179	39.87588	9.4100
17594554841883	08/01/2017 2:05:09	116.4216	39.86455	10.6901
17594554848724	08/01/2017 2:04:10	116.3882	40.1855	14.7602

The study area chosen in this study is a sub-network from Beijing which contains both expressways (the second ring road and third ring road) and local road segments (shown in Figure 1). The size of the area is 4.1 km by 4.6 km with 120 links in the road network. Given the road network and coordinates of floating car data, a map-matching algorithm from ArcGIS (30) has been applied to match all the vehicle information with the road link ID and then the spatiotemporal speed distribution can be obtained. The steps in map-matching process are shown in Figure 1.



### **Figure 3.1: Study Area and Map Matching Algorithm Examples**

After map-matching, data are aggregated into five-minute intervals. Since speed data typically fluctuate over shorter time intervals, the aggregation will make it smoother. Moreover, the time interval is updated in every 1 min with 5 min interval (e.g., the first time slice is 1min to 5min, the second time slice is 2 min to 6 min, the third time slice is 3 min to 7 min, and so on).

#### 3.2.2 Modeling Framework

In this study, to predict traffic speed at a network level, we adopt a computationally less expensive machine learning approach. In particular, we use XGBoost which is a scalable machine learning system based on gradient boosting (Chen and Guestrin, 2015). Gradient boosting, proposed by Friedman et al. (Friedman, 2001), merges a set of weak learners to a strong one, in an iterative fashion. XGBoost is a library optimized for boosting algorithm. The library provides a scalable, portable framework which has been used in many data science applications (Vanichrujee et al., 2018).

Boosting tree is a machine learning algorithm that uses linear combination of multiple weak learners to improve the learning accuracy of the algorithm. Regression tree is a base learner to solve regression problems (Vanichrujee et al., 2018). In order to constrain the number of leaf nodes and weights, regularization is added to the objective function of the Boosting Tree model in XGBoost algorithm. The regulation model prevents overfitting, resembling previous work on regularized greedy forest. But XGBoost simplifies the objective and algorithm for parallelization. By combining these insights, XGBoost can solve real-world problems at scale using a minimal amount of resources. More details of the XGBoost can be found in reference (Vanichrujee et al., 2018, Chen et al., 2016).



In recent years, XGBoost has been widely used in machine learning. One of the advantages of XGboost algorithm is its scalability, allowing it to operate much faster than existing machine learning algorithm in a single machine. It also scales to billions of examples in distributed or memory-limited settings. The scalability of XGBoost is due to several important systems and algorithmic optimizations such as a novel tree learning algorithm for handling sparse data and a theoretically justified weighted quantile sketch procedure that enables handling instance weights in approximate tree learning. Another feature of XGBoost is using column blocking which is for storing data in a compressed column format for parallel learning. Accordingly, parallel and distributed computing makes XGBoost learning faster which enables quicker model exploration (Alajali et al., 2018).

In the prediction model, the whole dataset is divided into a training dataset (80%) and a test dataset (20%), which are used to train the model and evaluate the performance of the model, respectively. To compare the performance of XGBoost model, liner regression (LR), decision tree regression (DTR), gradient boosting tree regression (GBTR), and random forest regression (RFR) models were chosen. In all models, the input variables were the previous 8 time slice link id, travel speed, service vehicle flow, and the time of the day, and the targets were the future 1-20 time slice travel speed of different links. The notations of different predictors and target are shown as follows:

$$\text{Previous Speed V: } \{v_{t-1}^{ln}, v_{t-2}^{ln}, v_{t-3}^{ln}, v_{t-4}^{ln}, v_{t-5}^{ln}, v_{t-6}^{ln}, v_{t-7}^{ln}, v_{t-8}^{ln}\}$$

where  $v_{t-p}^{ln}$  represents the average speed of link  $n$  in the interval  $p$  time slices before  $t$ .

$$\text{Previous Service Vehicle Flow D: } \{d_{t-1}^{ln}, d_{t-2}^{ln}, d_{t-3}^{ln}, d_{t-4}^{ln}, d_{t-5}^{ln}, d_{t-6}^{ln}, d_{t-7}^{ln}, d_{t-8}^{ln}\}$$

where  $d_{t-p}^{l_n}$  represents the average number of Didi vehicles traversing on link  $n$  in the interval  $p$  time slices before  $t$ .

Time of the day  $H$ :  $\{h_t\}$

where,  $h_t$  represents the hour of the time slice.

Link information  $L$ :  $\{l_1, l_2, l_3, \dots, l_n\}$

where,  $l_n$  represents a normalized variable created for link  $n$  with the range of  $l_n$  is  $\{0,1\}$ .

The targets of the model are:

The future travel speed  $V_p$  :  $\{v_p^{l_n,t}, v_p^{l_n,t+1}, v_p^{l_n,t+2}, \dots, v_p^{l_n,t+20}\}$

where,  $v_p^{l_n,t+h}$  represents the predicted average speed link  $n$  in the interval  $h$  time slices after  $t$  (i.e., prediction horizon=  $h$ ).

The evaluation metrics of the prediction performance are mean absolute error (MAE), mean absolute percentile error (MAPE) and root mean square error (RMSE):

$$MAE = \frac{1}{i * n} \sum_{n=1}^n \sum_{i=1}^i |v_r^{l_n,i} - v_p^{l_n,i}| \quad (1)$$

$$MAPE = \frac{1}{i * n} \sum_{n=1}^n \sum_{i=1}^i (|v_r^{l_n,i} - v_p^{l_n,i}| / v_r^{l_n,i}) \quad (2)$$

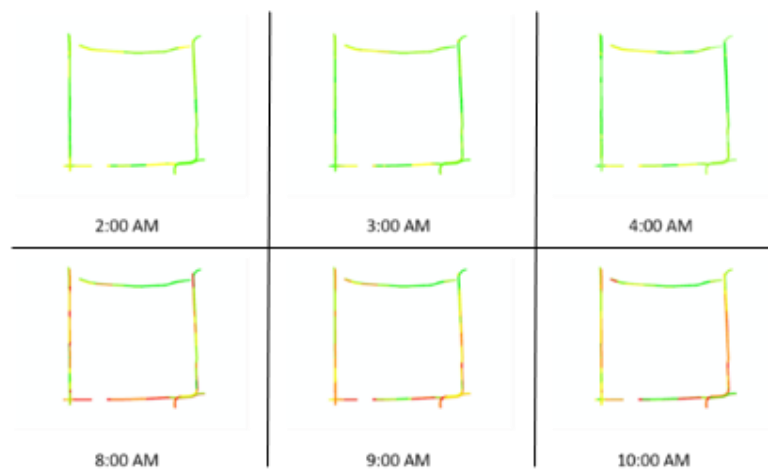
$$RMSE = \sqrt{\frac{1}{i * n} \sum_{n=1}^n \sum_{i=1}^i (v_r^{l_n,i} - v_p^{l_n,i})^2} \quad (3)$$

In the function,  $v_r^{l_n}$  represents the real travel speed of link  $n$ ;  $v_p^{l_n}$  represents the predicted travel speed of link  $n$ ;  $i$  represents the number of vehicles in different links;  $n$  represents the number of links.

### 3.3 Results

Figure 3.2 shows the travel speed of links in different time periods. In the figure, the travel speed is divided into 5 levels, represented by different colors (from red to green): Very congested: 6.24km/h - 24.98km/h; Congested: 24.98 km/h - 39.17 km/h; Regular: 39.17km/h - 51.43 km/h; Smoothed: 51.42 km/h - 62.23 km/h; Very smoothed: 62.23 km/h – 80.46 km/h. The time of the day affects travel speed since the peak hour has higher demand compared to other periods.

We observe that from 2:00 am to 4:00 am, traffic does not show any congestion, since most of the links show very smoothed level (green color). From 8:00 am to 10:00 am (the morning peak hour in Beijing), traffic experiences a congested situation, since many links show a very congested level (red color). Thus, the results indicate that it is essential to consider the time of the day as one of the elements when predicting the travel speed.



**Figure 3.2:** Traffic performance in different time periods

One of the challenging tasks to apply the machine learning methods in predicting the future traffic speed is identifying the best parameters. In this study, before applying the XGBoost methods as well as the benchmark methods, we applied grid-search to identify the optimal values

of parameters for each machine learning methods. The parameters of different methods are summarized as follows (the other values are default):

**Table 3.2:** Values of Parameters for Different Machine Learning Methods

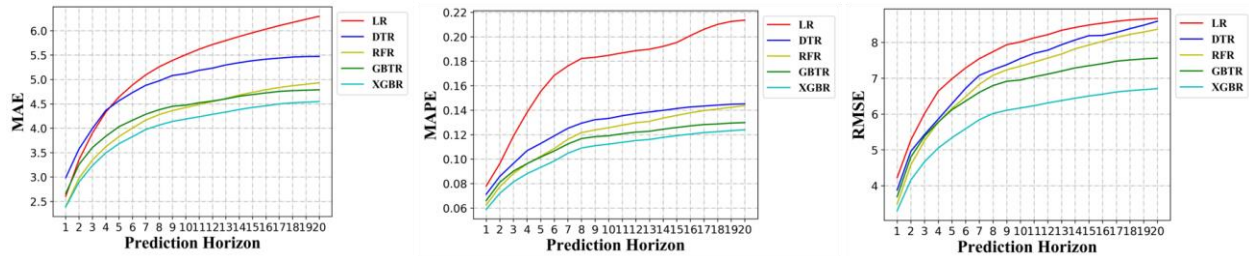
Methods	Max Depth	Learning Rate	Estimator
Linear Regression	N/A	N/A	N/A
Decision Tree Regression	N/A	N/A	N/A
Random Forest Regression	10	0.1	800
Gradient Boosting Tree Regression	10	0.1	800
<b>XGBoost Regression</b>	<b>15</b>	<b>0.3</b>	<b>800</b>

The results of XGBoost model and other machine learning approaches are presented in Figure 3.3. It shows that the XGBoost model can predict the short-term traffic speed of different links with high accuracy.

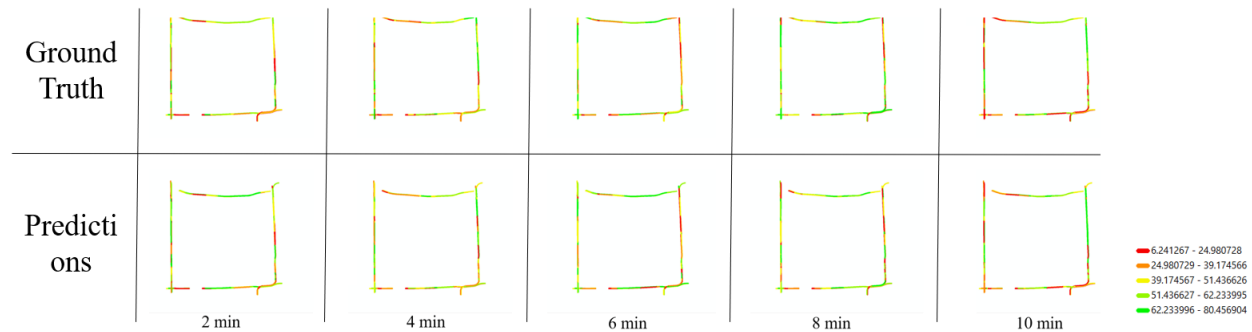
When compared with other methods, our results show that XGBoost achieves the highest prediction accuracies in all circumstances, followed by RFR and GBTR. For the 1 min prediction horizon, the MAE is 2.3861 km/h, and the corresponding MAPE is 0.0588, which means that the accuracy is more than 94%. When the prediction horizon is less than 7 time intervals, XGBoost can achieve more than 90% accuracy (MAPE is 9.84%), while all the other approaches have more than 10% errors, shown in Figure 3.3.

In addition, when predicting the travel speed of different prediction horizons, all the models show a similar pattern of decreasing prediction accuracy with increased prediction horizon. From Figure 3.3, it shows that the MAPE of XGBoost will increase from 0.059 (prediction horizon= 1 time interval) to 0.124 (prediction horizon= 20 time intervals). However, from Figure 3.3, for all the prediction horizon (from 1 to 20), XGBoost algorithm outperforms other machine learning methods, and it also has a good performance when forecasting for future 20 time slice travel speed

(12.4% MAPE). To intuitively show the performance of the prediction using XGBoost model, we compare the real traffic speed and predicted traffic speed in the selected network, showing in Figure 3.4. It presents that the colors (represent traffic speed level) of the predictions and ground truth are very similar, indicating that the XGBoost model works well.



**Figure 3.3:** Performance of Different Prediction Methods



**Figure 3.4:** Ground Truth vs. Predictions for different horizons using XGBoost

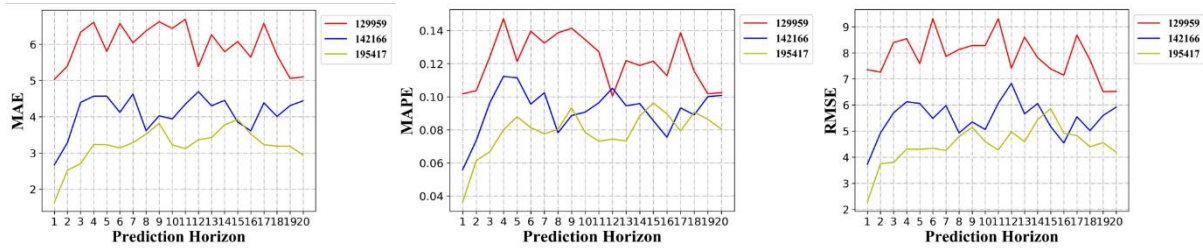
To reveal the efficiency of different prediction models, we investigated the training time of selected machine learning methods (see Table 3.3). We find that the linear regression and decision tree regression have the least training time (around 100 sec). Compared with random forest regression and gradient boosting tree regression, it presents that XGBoost regression is the most efficient model since it requires 913 seconds to train the dataset while the random forest

regression needs 1803s (nearly double) and gradient boosting tree regression needs 3458s (nearly 4 times). It indicates that XGBoost model has great potential applied in the traffic prediction.

**Table 3.3:** Training Time of Different Models

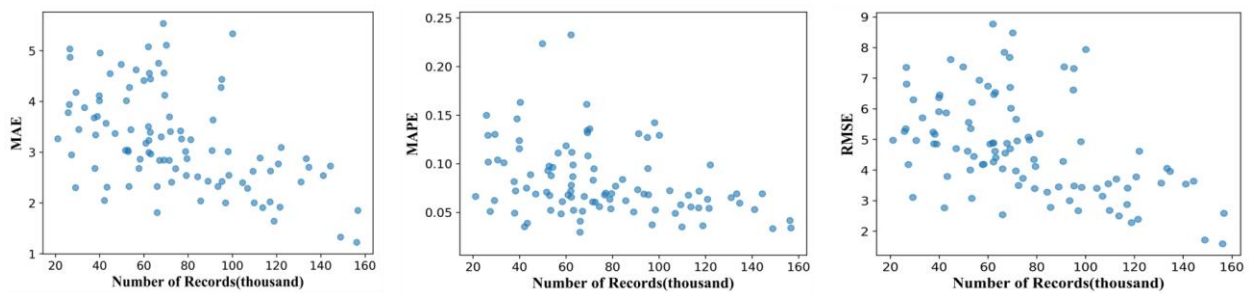
Models	Training Time (sec)
Linear Regression	104
Decision Tree Regression	100
Random Forest Regression	1803
Gradient Boosting Tree Regression	3458
<b>XGBoost Regression</b>	<b>913</b>

We observe that the accuracy of travel speed prediction using XGBoost differs across links. Figure 3.5 shows the prediction performance for three different links (link IDs: 129959, 142166 and 195417). It shows that the link 195417 has the best accuracy of prediction, since all the MAE, MAPE and RMSE of the link is the least, and the link 129959 presents the worst performance. To better understand the reason why the three links present different performance when predicting the future speed, the number of records of the three links are counted. It was found that the link 195417 has the most records which is 118,837, following by the link 142166 (74,343) and 129959 (26,473). Based on the number of records of the three links, it shows that the performance of the model is proportional to the number of records from the three links. Thus, the accuracy of the model appears to be affected by the number of records of different links.



**Figure 3.5:** Performance of Prediction Methods of Different Links

To validate this assumption, the prediction accuracy (prediction horizon= 5 time intervals) and number of service vehicles of all the links were given, and their relationship was further examined in Figure 3.6. The relationship indicates that the MAE and RMSE are inversely proportional to the number of records and the MAPE also decreases when the number of records increases. Based on the results, it can be observed that the number of service vehicles will significantly affect the accuracy of the prediction model, since the pattern will be more stable when the dataset is larger. Thus, when predicting the future traffic speed which requires high accuracy, it reveals that we can not only focus on the prediction model, such as machine learning model and deep learning model, but we should also account for the data collection technologies and the quality and stability of the dataset.



**Figure 3.6:** Prediction performance vs. the number of records



### 3.4 Discussion

In this section, we present a machine learning approach XGBoost to predict the travel speed of urban links, which will be more beneficial for future intelligent transportation systems and smart city applications. Previous works have mainly focused on the speed data collected from traffic sensors, with issues such as data sparsity and low coverage. In this study, taxicab data—with higher quality and larger coverage—are used for prediction purpose. The selected study region is a sub network in Beijing with 120 links containing both expressway and local roads. Four other machine learning models (linear regression, decision tree regression, random forest regression and gradient boosting tree regression) are also tested to compare with the XGBoost, based on prediction performance using metrics such as MAE, MAPE, and RMSE. However, providing relatively long-term prediction is also meaningful for not only recommending the best route but also for transportation management. Thus, in this study, the proposed models are applied for speed prediction (prediction horizon= 1 to 20 time intervals). The results show that XGBoost outperforms other machine learning approaches for all prediction horizons.

We also find that, for the same algorithm, different links have different performance levels when predicting the future travel speed. Three links have been found to have different accuracy with XGBoost model, with their accuracy levels are proportional to the number of records observed. To validate the assumption that the number of records observed in a link will affect the performance of the model, the relationship between the evaluation metrics (MAE, MAPE, and RMSE) and the number of records of all links is further examined. The results show that the number of records affect the performance of the prediction model and all the three-evaluation index indicates a proportional relationship between the accuracy and the number of records. This finding reveals that in addition to focusing on the prediction model—such as machine

learning/deep learning models—we should also consider different data collection technologies to obtain a high-quality larger dataset.

Future work can be conducted by adding spatial, temporal, service flow information into a deep learning model. This implies that traffic speed from adjacent regions can be added as additional inputs. Furthermore, prediction performance over different data aggregation levels (both spatial and temporal) should be investigated.

## CHAPTER 4: MOBILITY BEHAVIOR ANALYSIS

### 4.1 Introduction

Spatio-temporal patterns of human mobility gives information on how a city functions. Understanding individual mobility behavior, from different perspectives, is important to solve many city problems such as urban planning (Sun et al., 2016, Tian et al., 2010), traffic management (Chen et al., 2016), public safety (Horanont et al., 2013, Lu et al., 2012), intelligent transportation system (Zhang et al., 2011), smart cities (Pan et al., 2013), public transportation (Zhao et al., 2018), disease spread and control (Bajardi et al., 2011, Wesolowski et al., 2012) and emerging issues such as autonomous vehicle operations (Bansal and Kockelman, 2017) and mobility as a service design (Jittrapirom et al., 2017). In recent years, a wide range of emerging human movement data sources—such as bank notes (Brockmann et al., 2006), social media data (Hasan et al., 2013b, Jurdak et al., 2015), mobile phone call details records (Huang et al., 2018), smart card transactions (Zhao et al., 2018), and floating car observations (Chen et al., 2019, Peng et al., 2012, Veloso et al., 2011, Zheng et al., 2018)—have been used to uncover individual mobility behavior (Gonzalez et al., 2008) and commuting patterns (Ma et al., 2017b). In this section, we report the findings on mobility behavior of a new population group—users of emerging on-demand ride-hailing services—after analyzing large-scale trip data from a ride-hailing platform.

With the emergence of ride-hailing services such as Uber, Lyft and Didi, massive passenger movement data from these platforms have a tremendous potential to reveal individual travel behavior patterns as well as commuting activity characteristics. In this study, we analyze the spatio-temporal patterns of individual mobility using the movement data extracted from Didi, a Chinese ride-hailing service. First, we propose a distance-based algorithm to identify the visited places of different passengers. Second, given the visited places of passengers, we investigate the

spatio-temporal patterns of individual movements. Then, for every individual user, we detect their home and workplace based on the probability of visiting different places at different time periods (morning and evening peak hours). Finally, we reveal individual mobility patterns when using ride-hailing services from different perspectives such as trip generation, gap time, number of visited places and their rank, spatial distribution of home and work place, travel distance and travel time. The resulting distributions show the potential of modeling the generation mechanism of ride-hailing service demand. Such models will enable high-fidelity (e.g., at individual level) simulation of demand prediction, dispatching, ride sharing, and pricing applications of ride-hailing services. To the best of our knowledge, this is the first study that reveals individual mobility patterns of ride-hailing service users based on large-scale data from a ride-hailing platform.

## **4.2 Data and Methods**

### **4.2.1 Dataset Description**

In this study, we have collected a large-scale dataset from Didi (the largest Chinese ride-hailing service in Beijing) for analysis. Beijing is a metropolis in northern China and the capital of the People's Republic of China (PRC). Nowadays, the number of vehicles has reached to approximate to 6 million and more than 27 million trips are made every day in Beijing. The study region covers the area inside Beijing's 6<sup>th</sup> ring road, which is the core region of Beijing. The dataset used in this study was extracted from the ride-hailing service platform Didi from May 1, 2017 to June 31, 2017. The dataset records more than 3 million Didi platform users with around 20 million trips. It includes 11 fields as listed in Table 1 with detailed data attributes.

**Table 4.1: Detailed Data Attributes**

Fields	Field Name	Field Type	Field Description
R_id	Record ID	String	The record id of one trip
P_id	Passenger ID	String	The passenger id of one trip
D_id	Driver ID	String	The driver id of on trip
O <sub>LNG</sub>	Longitude of Origin	Floating	The longitude of the origin
O <sub>LAT</sub>	Latitude of Origin	Floating	The latitude of the origin
D <sub>LNG</sub>	Longitude of Destination	Floating	The longitude of the destination
D <sub>LAT</sub>	Latitude of Destination	Floating	The latitude of the destination
O <sub>Time</sub>	Start Time	Timestamp	The timestamp of the origin
D <sub>Time</sub>	Arrive Time	Timestamp	The timestamp of the destination
L	Travel Distance	Floating	The travel distance of the trip
C	Cost	Floating	The price of the trip record

Raw movement data from a ride-hailing platform has several issues. For example, GPS errors may be caused by either blockage of the GPS signal or hardware/software bugs during the data collection process. To clean the raw data, we followed the following four steps:

**Step1:** Convert the current coordinate system of Didi’s data to the Worldwide Geodetic System 1984 (WGS84) coordinate system;

**Step2:** Remove the data which have the coordinates (origin or destination) outside Beijing’s 5th ring road;

**Step3:** Remove the data which contains invalid values due to GPS errors;

**Step4:** Based on the limited speed, remove the trips with average travel speed ( $Distance / (D_{Time} - O_{Time})$ ) above 120 km/h.

#### 4.2.2 Distance-Based Visited Place Generation Algorithm

Passenger movement data provide the GPS coordinates of origins and destinations. However, two origins or destinations can belong to the same place (e.g. home or workplace) with different coordinates possibly due to GPS errors or different boarding points from the same location. Based on individual travel patterns, their home and workplace can be inferred. Passengers typically leave

home and arrive at workplace in the morning peak hour and have a reverse travel direction in the evening peak hours. In this study, to identify the visited places of different individual users and the function (home or workplace) of these places, we have applied several heuristic rules as follows:

**Rule1:** For each individual user, if the distance between two locations (origins or destinations) is less than 500m, then these two locations are defined as the same place.

**Rule2:** For each individual user, for each location, we count the number of origins in the morning peak hour (6:00 am – 11:00 am) and the number of destinations in the evening peak hours (15:00 pm – 20:00 pm). Then, define the location which has the largest sum as the home place.

**Rule3:** For each individual user, we count the sum of the number of destinations in the morning peak hours (6:00 am – 11:00 am) and the number of origins in the evening peak hours (15:00 pm – 20:00 pm) for each location. Then, define the location which has the largest sum as the workplace. Workplaces may be schools for students, may be daily routine places for some users, and particularly locations, including shopping, social or recreational functions. However, they will all be referred as "workplaces" in this study.

According to **Rule1**, we developed a distance-based visited place generation algorithm (*DBVPGA*) to identify the visited places of each individual user. According to the distance between different coordinates, the *DBVPGA* will detect whether the origin or destination is a new place and then assign an ID, as a visited place (start from 0), to the origin and/or destination. The key definitions of *DBVPGA* are shown as follows:

- $d_{th}$ : the threshold distance used to identify places ( $d_{th} = 500$  meters).
- $d_{i,j}$ : the distance between point  $i$  and point  $j$ .
- $O_n$ : the  $n^{th}$  origin point of a user.

- $D_n$ : the  $n^{th}$  destination point of a user.
- $MTN$ : the number of trips per month made by a user.
- $PID$ : the ID of visited places of a user (0, 1, ..., n).
- $maxPID$ : the maximum PID of a user.
- $VPF$ : the types of visited places of a user (0-home, 1-work, 2-other).

We develop an algorithm to convert the coordinates of the origins and destinations into relative IDs of visited places for an individual ( $PIDs$ ). Briefly speaking, for each individual user, we have a list of coordinates of origins and destinations of all the trips made by the user. Every origin and destination are defined as different points and we start from the origin of the first trip by setting it as the first visited place ( $PID = 0$ ). Then, choose the destination of that trip as the second point and compare the distance between this point and the previous point ( $d_{i,j}$ ) with the threshold distance ( $d_{th}$ ). If  $d_{i,j}$  is more than  $d_{th}$ , then set the second point as a new visited place, the  $PID$  of the second place is 1 and the  $maxPID$  is added by 1. In this way, two different visited places are generated. If  $d_{i,j}$  is less than  $d_{th}$ , then the second point is the same visited place as the first point, and the  $PID$  of the second place is also 0. Likewise, for the other points, we compare the distance between them and existing visited places with  $d_{th}$ , if all the distances  $d_{i,j}$  are more than  $d_{th}$ , then generate a new visited place and the  $maxPID$  is added by 1. Otherwise, if the distance between the point and any existing place is less than  $d_{th}$ , then the  $PID$  of this point will be the same as the specific existing place. We iterate over all the points until every point has its own  $PID$ . The algorithm is described as follows:

---

**Algorithm 1** Distance-Based Visited Place Generation Algorithm

---

**Input:** Passenger ID, the list of coordinates of origins and destinations, number of trips per month (MTN)

**Output:** PID

---

1. For each individual user:
2.      $PID_{oi} = 0, \max PID = 0.$
3.     For  $i$  from 2 to MTN:
4.          $ON = i - 1$
5.         For  $j$  from 1 to ON:
6.             If  $l_{oi,oj} < l_{th}$ :
7.                  $PID_{oi} = PID_{oj}$
8.             End if
9.         End for
10.         If  $PID_{oi} = N/A$ :
11.              $\max PID = \max PID + 1$
12.              $PID_{oi} = \max PID$
13.         End if
14.     End for
15.     For  $i$  from 1 to MTN:
16.         If  $l_{di,oi} < l_{th}$ :
17.              $PID_{di} = PID_{oi}$
18.         End if
19.          $DN = i - 1$
20.         For  $j$  from 1 to DN:
21.             If  $l_{di,dj} < l_{th}$ :
22.                  $PID_{di} = PID_{dj}$
23.         End for
24.         If  $PID_{di} = N/A$ :
25.              $\max PID = \max PID + 1$
26.              $PID_{di} = \max PID$
27.         End if
28.     End for
29. End for

---

After running the algorithm, the visited places of each user are generated, the function (home place or workplace) of the visiting places can be identified according to **Rule 2** and **Rule 3**.

### 4.3 Empirical Results

#### 4.3.1 Temporal Pattern - Trip Generation

Based on the origin and destination information, a trip can be characterized by its travel purpose.

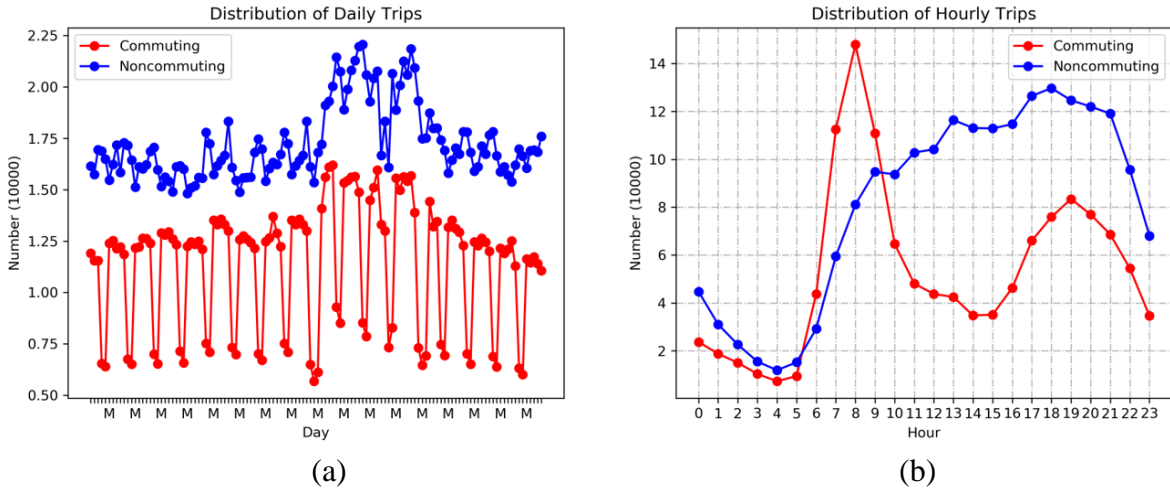
In this study, we analyze the individual trip generation patterns by decomposing the on-demand



ride-hailing service trips into two groups – commuting and non-commuting trips according to their travel locations. We define a trip which is made between the residence and the work place of a user as a commuting trip. When a trip contains at least one location which is neither the residence nor the work place, we define it as a non-commuting trip.

Figure 4.1a shows the daily trip distribution of commuting and non-commuting trips of all the selected users of the on-demand ride-hailing service. It reveals weekly periodicity of passengers' travel behavior. Most of the on-demand service users tend to travel more frequently on weekdays compared to weekends and festivals (May 1st, the Labor Day, is a holiday in China). However, the periodicity characteristics for the commuting trips and non-commuting trips shows significant difference. In terms of commuting trips, since people always work on weekday, the number of trip generation of commuting trips decrease sharply on weekends (but not 0). The demand of non-commuting trips is more than that of commuting trips. The weekly periodicity indicates that the demand on weekends of non-commuting trips will increase smoothly. Because the decrease demand for commuting trips is more than the increase demand for non-commuting trips in general, the whole demand for all trips is higher on weekday compared to weekends.

Figure 4.1b presents the hourly distribution of trips, indicating that the on-demand ride-hailing service trips for commuting have a typically bimodal distribution while the distribution for non-commuting trips is unimodal. For commuting trips, peak demand is seen from 7 am to 9 am (morning peak hour) and from 5 pm to 9 pm (afternoon peak hour). For non-commuting trips, peak demand is seen from 5 pm to 9 pm. The highest demand of ride-hailing service is seen around 8 am for commuting purpose. From 10 am to midnight, the demand of non-commuting trips exceeds the demand of commuting trips.



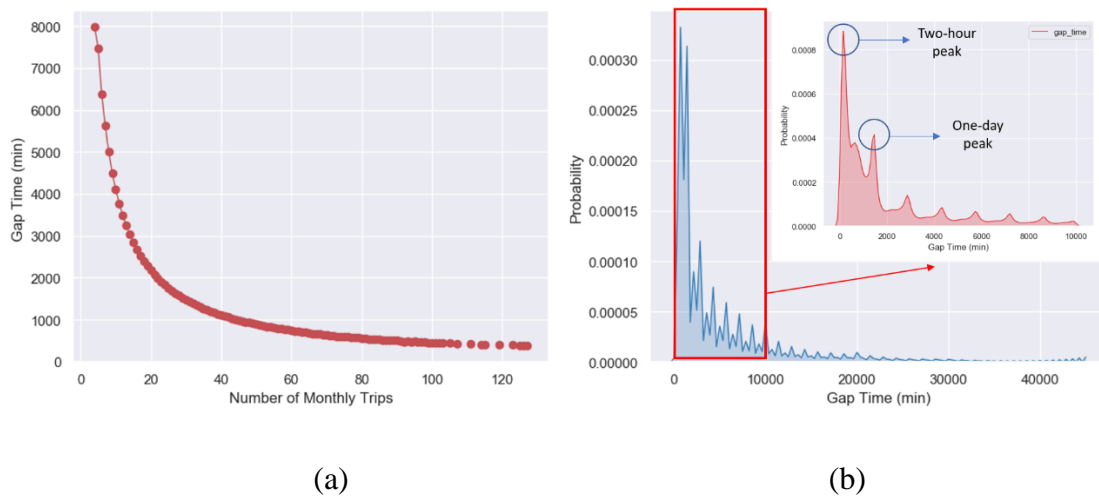
**Figure 4.1:** The probability of the demand of ride-hailing service (March 1-June 30, 2017): (a) the distribution of the number of daily trips using the ride-hailing service (‘M’ means Monday of every week) (b) the distribution of average hourly trips number of ride-hailing service.

#### 4.3.2 Gap Time

One of the most important characteristics of a ride-hailing service is how long does it take for a user to make the next trip. To uncover the distribution of the time spent between to make a new trip, we create a variable called the gap time. It is defined as time difference (start time of trips) between two consecutive trips for a user. To determine this distribution, we select all the users who have made at least 2 trips in a month.

Figure 4.2a shows the distribution of average gap time of user groups with the number of trips made in a month. The results indicate that with the increase of monthly trips, the gap time between consecutive trips decreases sharply for the beginning and then the declination trend of gap time turns to slower after the number of monthly trips is more than 20. The maximum gap time is found as around 8000 min (5.6 days) when users have only 2 monthly trips. When the number of monthly trips is more than 80, the gap time is close to the minimum value of around 300 min (5 hours).

Figure 4.2b presents the distribution of gap time of all the users in the observation period (one month). There are two distinct peaks which happens with two-hour and one-day gaps, respectively. This finding is different from the current mobility research (16) which shows a 9-hour peak when it comes to the public transport users. Additionally, the distribution also indicates that multi-day local peaks, which might be closely related to the passengers' commuting behavior or regularity patterns in requesting rides.



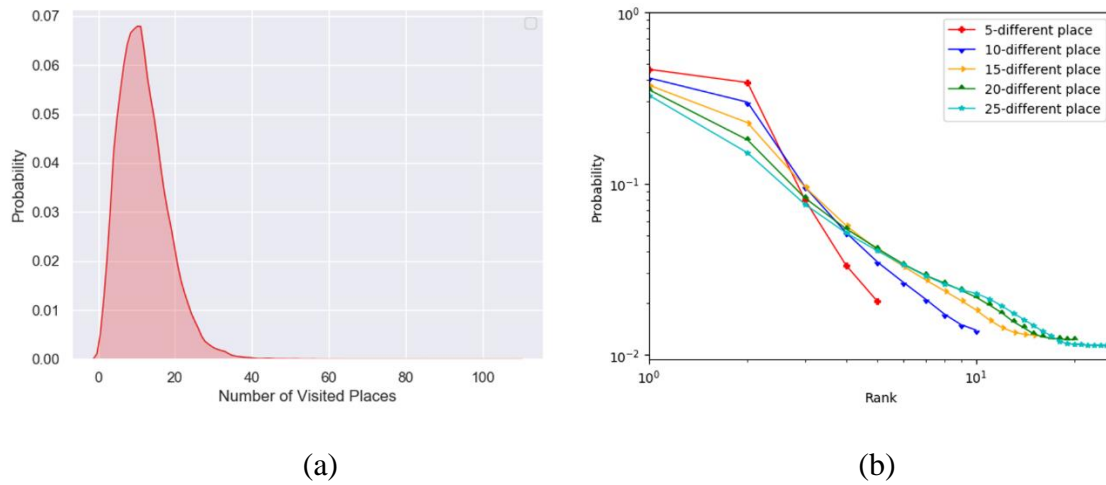
**Figure 4.2:** The distribution of gap time. (a) the average gap time (min) vs. their corresponding users' groups; (b) the probability of different gap time (min) in two scales

#### 4.3.3 Number of Visited Places and their Rank

Each ride-hailing service user visits a specific number of different locations within the observation period. We rank the visited places based on the frequency of visits to those places and determine the probability of each visited place. For instance, for a user, a visited place with rank 1 means the most visited location, a visited place with rank 2 represents the second-most visited location for the user, and so on.

The number of visited places and the rank of those locations play an essential role for mobility pattern analysis. To uncover the distribution of the number of visited places and the rank of visited places, we select the passengers who have at least 30 monthly trips so that enough trips are generated to reveal the spatial patterns. These passengers are defined as frequent users hereafter in the study and we randomly choose 40,000 frequent users in this study.

Figure 4.3a presents the distribution of the number of visited places—indicating that majority of the frequent users of the ride-hailing service visit on average 8 to 12 different places in a month. Given that a frequent user makes at least 30 monthly trips and every trip contains two places (origin and destination), he/she has a high probability to visit the same places when using the service. Additionally, it indicates that users tend to spend distinct time on different visited locations. Thus, to uncover the regularity patterns, the probability distribution of visiting a place over the rank of the visited place is presented in log-log scale in Figure 4.3b. From the distribution, we can find that most of the users' trips are concentrated in a few locations, especially the first rank visited place. For instance, users who visited 5 different places, the most visited place accounts for nearly 50% of the total trips. When a user visits more places, the probability of the most visited place slightly declines. Additionally, the probability of the second most visited place (i.e., rank =2) is close to the most visited place when the number of visited places is low. Users who have visited more places, the difference between the probability of the first two rank visited places becomes larger. It shows that the distribution of the visited place rank follows a Zipf's law when the number of visited places is high.



**Figure 4.3:** The distribution of passengers’ visited place number and the probability of visited place rank: (a) the probability of different visited place number. (b) the probability of visiting different places based on its rank level in log-log scale

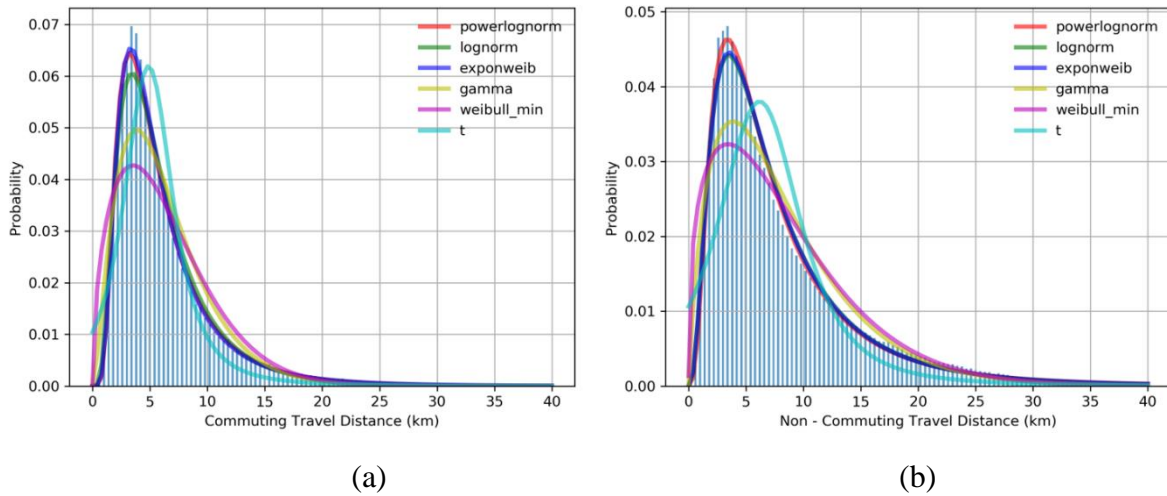
#### 4.3.4 Distribution of Travel Distance

To reveal the spatial patterns of individual mobility under ride-hailing service, the distributions of travel distance for both commuting trips and non-commuting trips are analyzed. In the existing studies, the smart card data and smart phone data, which are commonly used data source for individual mobility patterns, which can provide the displacement information of human movement. However, movement data extracted from a ride-hailing service platform offer us high-precision travel distance data rather than displacement data (Wang et al., 2015), since the location of each trip’s origin and destination can be accurately determined. Additionally, the commuting (home to work trips) patterns and non-commuting trips’ patterns are significant aspects in human travel behavior studies, which are seldom investigated due to data source limitation. Thus, to uncover the individual commuting and non-commuting travel behavior, we choose the frequent users (more than 10 monthly trips). In addition, we also fit the travel distance distribution by 6

commonly used distributions mentioned before, which will be beneficial for establishing generation mechanism in traffic simulation.

Figure 4.4 shows the distribution of travel distance for commuting trips and non-commuting trips. It can be found from the figure that the travel distance for non-commuting trips is more than travel distance for commuting trips. The peaks of the two distributions occur at around 5 km. The average travel distance of commuting trips and non-commuting trips are respectively 6.298 km and 8.467 km. In terms of long distance trips, it indicates that the commuting trips which have a more than 15 km travel distance accounts for 5.58% from our dataset and the ratio for the long-distance trips ( $> 15$  km) for non-commuting trips is 14.62%, which is identical with our daily experience that person do not prefer to have a long commuting trip by taxicab or on-demand ride-hailing service and literature (Wang et al., 2015) show similar results.

To capture the on-demand ride-hailing travel distance distribution, we use 6 existing statistical distributions - log-normal, Weibull, gamma, student's t, exponentiated Weibull and power log-normal – to fit the travel distance distribution with a K – S test to evaluate the performance. Table 4.2 presents the results of K-S test for travel distance. From the results, we can find that power log-normal distribution fit best for both the commuting travel distance and non-travel distance, since the power log-normal distribution have the lower D value (0.235 for commuting trips and 0.092 for non-commuting trips) and higher p-value (0.104 for commuting trips and 0.977 for non-commuting trips), which indicates a higher probability to accept the null hypothesis that the two distributions come from one identical distribution.



**Figure 4.4:** The distributions of travel distance per trip with fitting curves of selected distributions: (a) distribution of travel distance (km) per trip for commuting trips; (b) distribution of travel distance (km) per trip for non-commuting trips

**Table 4.2:** The results of K-S test of selected distribution for travel distance

Distribution K-S Test	Commuting Distance			Non - commuting Distance		
	D	<i>p</i> -value	Parameters	D	<i>p</i> -value	Parameters
Power log-normal	<b>0.235</b>	<b>0.104</b>	$p = 5.64; \sigma = 1.00$	<b>0.092</b>	<b>0.977</b>	$p = 0.95; \sigma = 0.74$
Log-normal	0.255	0.062	$\mu = 1.52; \sigma = 0.61$	0.112	0.875	$\mu = 1.85; \sigma = 0.75$
Exponential Weibull	0.255	0.062	$k = 6.01; \alpha = 0.89;$ $\lambda = 2.54$	0.112	0.875	$k = 13.27; \alpha = 0.50;$ $\lambda = 0.72$
Gamma	0.608	0.000	$\alpha = 1.04; \beta = 1.85$	0.176	0.377	$\alpha = 1.88; \beta = 4.39$
Weibull	0.314	0.010	$k = 1.53; \lambda = 7.02$	0.275	0.036	$k = 0.80; \lambda = 3.53$
Student's t	0.355	0.002	$\nu = 2.13$	0.235	0.104	$\nu = 2.32$

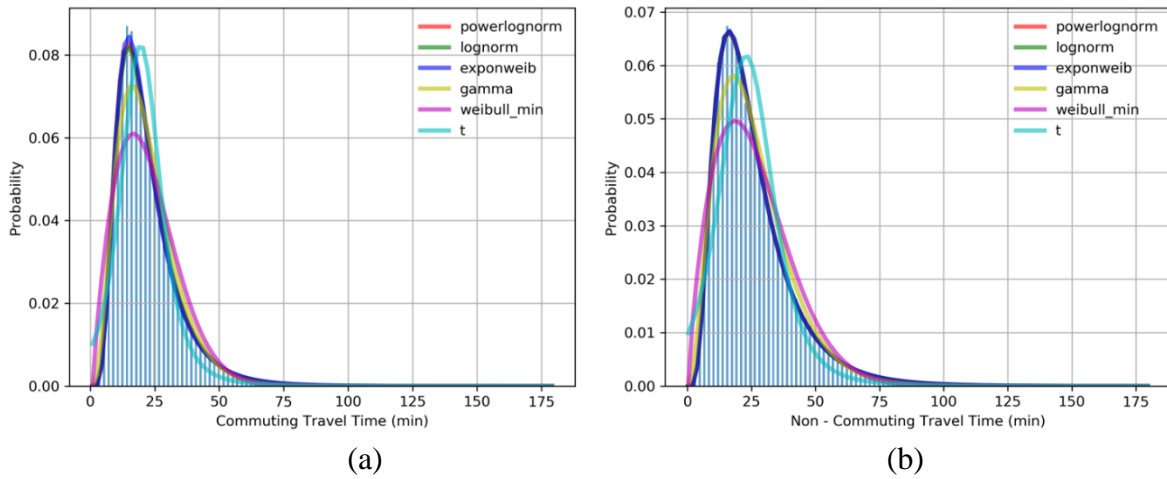
#### 4.3.5 Distribution of Travel Distance

Besides the travel distance, the travel time of each on-demand ride-hailing service also play an important role in spatial to reveal the spatial regularity, which reflect the time cost of trips. Figure

4.5 presents the distribution of travel time for both commuting trips and non-commuting trips. From the results, it indicates that the peak of the commuting travel time occurs at around 12 min and the peak of the non-commuting travel time happen at around 14 min. The average travel time for commuting trips and non-commuting trips are respectively 22 min and 26 min. In general, the travel time for non-commuting trips is more than that for commuting trips, which shows a similar result with the distribution of travel distance.

In addition, we also use 6 commonly used distribution to fit the travel time distribution, and the results can be seen as table 4.3. The results of K-S test show that the commuting travel time distributions fit best to the Exponential Weibull distribution (with  $D = 0.157$  and  $p$ -value  $= 0.527$ ) and the non-commuting travel time distributions fit best to the log-normal distribution (with  $D = 0.058$  and  $p$ -value  $= 1.000$ ).





**Figure 4.5:** The distribution of travel time per trip with fitting curves of selected distributions:  
 (a) Distribution of travel time per trip for commuting trips. (b) distribution of travel time per trip  
 for non-commuting trips

**Table 4.3:** The results of K-S test of selected distribution for travel time

Distribution K-S Test	Commuting Travel Time			Non - commuting Travel Time		
	D	<i>p</i> -value	Parameters	D	<i>p</i> -value	Parameters
Power log-normal	0.181	0.347	$p = 0.12; \sigma = 0.23$	0.081	0.995	$p = 0.15; \sigma = 0.22$
Log-normal	0.360	0.002	$\mu = 0.62; \sigma = 1.52$	<b>0.058</b>	<b>1.000</b>	$\mu = 3.16; \sigma = 0.52$
Exponential Weibull	<b>0.157</b>	<b>0.527</b>	$k = 11.00; \alpha = 0.77;$ $\lambda = 4.93$	0.083	0.993	$k = 15.29; \alpha = 0.63;$ $\lambda = 3.53$
Gamma	0.412	0.000	$\alpha = 3.58; \beta = 5.89$	0.294	0.020	$\alpha = 3.19; \beta = 8.02$
Weibull	0.471	0.000	$k = 1.86; \lambda = 23.92$	0.255	0.062	$k = 1.64; \lambda = 31.40$
Student's t	0.200	0.236	$v = 3.21$	0.093	0.976	$v = 3.69$

#### 4.4 Discussion

In this study, we have analyzed large-scale trip data extracted from a ride-hailing service platform (Didi) in China to understand individual mobility patterns. To the best of our knowledge, this is the first study reporting the mobility behavior of ride-hailing service passengers. Human mobility can be interpreted as the movement patterns in a spatio-temporal scale. To uncover the spatio-temporal patterns of individual movements, we have analyzed the distribution of the trip generation, gap time, number and rank of visited places, travel distance, and travel time. In addition, to capture the patterns of commuting behavior, we characterize the trips into two groups: commuting and non-commuting trips according to the travel purpose of individuals.

For temporal patterns, first, the distributions of daily and hourly trips reveal the travel regularities of ride-hailing service users. It indicates that people tend to use more on-demand service on weekdays and the trips on weekends are about 20% less than the trips on weekdays. Additionally, the patterns of hourly trip generation distributions show differences between commuting trips and non-commuting trips. The distributions of trip time for commuting trips reveal a bimodal distribution and a unimodal distribution for non-commuting trips. The morning peak hours of non-commuting trips vanish because most of the trips of non-commuting trips are for leisure or entertainment activities. The findings deviate from that of the previous study on temporal trip generation patterns. In the study (Ma et al., 2017b), it indicates that both of the distribution of commuters and non-commuters using public transit follow bimodal distributions and the total trips of commuters are significantly more than that of non-commuters, which presents a difference with our results. The difference shows that compared with the public transit, the on-demand ride-service shows less attraction for commuters, especially in morning peak hours.

Another important aspect reported in this section is the distribution of gap time. In recent years, most studies have analyzed the waiting time or stay time patterns based on mobile phone data (Gonzalez et al., 2008) and interval distribution of taxi trajectory from drivers' perspective (Veloso et al., 2011, Zheng et al., 2018). To fill the gap that no study has investigated the distribution of the time interval between two consecutive trips of ride-hailing service users, we have analyzed the average gap time (time interval between two consecutive trips) distributions of on-demand service users from two aspects. First, the amount of gap time between consecutive trips is inversely proportional with the number of monthly trips. Second, the distribution of gap time follows a log-normal distribution with local spikes, which presents a similar pattern with the stay time by smart card data (Hasan et al., 2013a) and the return time based on mobile phone data (Gonzalez et al., 2008). In addition, previous research (Gonzalez et al., 2008) also found the gap time probability is characterized by several local peaks at 24h, 48h, 72h and so on, which shows a strong temporal regularity to make the next trip and captures the temporal periodicity inherent to human mobility. However, from the ride-hailing service users' perspective, the gap time probability has a two-hour local peak compared to the 9-hour local peak in the previous studies (Hasan et al., 2013a). It probably indicates that people tend to use public transportation for commuting trips which have a 9-hour local peak, while people prefer to use on-demand services for leisure or other activities which have a two-hour local peak.

To find the spatial regularities of individual mobility, we identify each user's visited places and rank those places according to the number of times a user has visited a place. From the results, we find that frequent users tend to have 8-10 different visited locations and visit more the top two locations. This shows a great regularity pattern of the on-demand service users when they make a trip. Under most circumstances, the most visited place is home or work place. Previous research

found similar results with the data extracted from smart card transactions (Hasan et al., 2013a), mobile phone call records (Gonzalez et al., 2008) and taxi trajectories (Peng et al., 2012). Studies (Peng et al., 2012) reveal that the probability of visiting locations follows a Zipf's law. Hasan et al. (Hasan et al., 2013a), using smart card data, found that the two most visited places have similar probabilities and the probability distribution of the places with rank greater than 2 follows a Zipf's law. When it comes to ride-hailing service users, we find the similarity of the probabilities of the two most visited places, similar to the results found by Hasan et al. (Hasan et al., 2013a). Additionally, the distribution of the visited place rank probabilities seems to follow a Zipf's law when the number of visited place is higher.

Additionally, we visualize the spatial distribution of home and work places of on-demand ride-hailing service users. Compared with the public transit users (Ma et al., 2017b), jobs-housings of on-demand ride-hailing users does not show severe imbalance possibly because people are less likely to take ride-hailing services for long-distance commuting. We also present the travel distance distribution in spatial scale for commuting and non-commuting trips to validate that on-demand users prefer to commute in short distance.

Finally, we have reported the distributions of travel distance and travel time both for commuting trips and non-commuting trips of ride-hailing service users, capturing the patterns of another significant aspect of spatial regularity. The travel distance and travel time for both commuting trips and non-commuting trips show similar patterns observed in other studies (Zheng et al., 2018, Zhao et al., 2015). It is worth mentioning that the average travel distance and travel time distribution of commuting trips presents more left-skewed compared to that of non-commuting. It implies that people tend to travel less distance when they commute by a ride-hailing

service, probably due to economic considerations. Both commuting trips and non-commuting trips' travel distance follow a power log-normal distribution from our results which is different from the previous studies (Zhao et al., 2015). In addition, we also fit the distribution of travel time for commuting trips and non-commuting trips which is always ignored in the existing studies - the travel time distribution of commuting trips follows an exponential Weibull distribution and the travel time distribution of non-commuting trips follow a log-normal distribution.

#### **4.5 Implications**

The results of this study will provide several implications for policy makers, traffic management and urban planners, which are summarized as follows:

1. Urban land use characteristics have significant influence on individual commuting patterns(Suzuki and Lee, 2012). Thus, understanding the spatial distributions of individual residences and work places are the key points for urban planning and policy decisions (Aguilera and Voisin, 2014). In this study, we provide a cost-effective way to collect large-scale data on individual home and work places based on the emerging ride-hailing trip data.

2. Predicting traffic states is one of the critical aspects of deploying intelligent transportation systems. Accurate and reliable prediction of travel states such as travel time enables people to make informed travel decisions. The individual mobility patterns presented in this study will be beneficial to establish high-resolution demand prediction model in individual level, which can be applied to improve the performance of transportation network (Wen et al., 2019).

3. The distribution of trip generation, gap time, travel distance and travel time can be used in agent-based traffic simulations, which is one of the most essential tools for evaluating the

expected performance of a new policy as well as innovative technologies, such as the connected and autonomous vehicles and mobility as a service (Wen et al., 2019).

4. The development of ride sharing would be one of the effective measurements to mitigate the congestion problems and meet the commute demand, especially with autonomous vehicles since it can expend automobile occupancy (Lavieri and Bhat, 2019). The individual mobility behaviors will provide indispensable experiences to develop matching algorithms for potential passengers and drivers, which can significantly improve the service of ride sharing.

## **CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS**

### **5.1 Summary**

In the first problem, we present a machine learning approach XGBoost to predict the travel speed of urban links, which will be more beneficial for future intelligent transportation systems and smart city applications. In this study, taxicab data—with higher quality and larger coverage—are used for prediction purpose. The selected study region is a sub network in Beijing with 120 links containing both expressway and local roads. Four other machine learning models (linear regression, decision tree regression, random forest regression and gradient boosting tree regression) are also tested to compare with the XGBoost, based on prediction performance using metrics such as MAE, MAPE, and RMSE. However, providing relatively long-term prediction is also meaningful for not only recommending the best route but also for transportation management. Thus, in this study, the proposed models are applied for speed prediction (prediction horizon= 1 to 20 time intervals). The results show that XGBoost outperforms other machine learning approaches for all prediction horizons.

We also find that, for the same algorithm, different links have different performance levels when predicting the future travel speed. Three links have been found to have different accuracy with XGBoost model, with their accuracy levels are proportional to the number of records observed. To validate the assumption that the number of records observed in a link will affect the performance of the model, the relationship between the evaluation metrics (MAE, MAPE, and RMSE) and the number of records of all links is further examined. The results show that the number of records affect the performance of the prediction model and all the three-evaluation index indicates a proportional relationship between the accuracy and the number of records. This finding reveals that in addition to focusing on the prediction model—such as machine

learning/deep learning models—we should also consider different data collection technologies to obtain a high-quality larger dataset.

In the second problem, we reveal the spatio-temporal patterns of individual mobility from the perspective of ride-hailing service passengers. The empirical analysis of massive movement data provides us deeper insights on individual mobility patterns at a city level. Regarding temporal movement patterns, we capture the difference of trip generation characteristics between weekdays and weekends and the distribution of gap time between consecutive trips. In terms of spatial mobility patterns, we visualize the distribution of home and workplaces as well as the travel distance in spatial scale and observe the distribution of the number of visited place and their rank and report the distribution of travel distance and travel time.

## **5.2 Limitations and Future Research Direction**

In the first problem, future work for the traffic status prediction can be conducted by adding spatial, temporal, service flow information into a deep learning model. This implies that traffic speed from adjacent regions can be added as additional inputs. Furthermore, prediction performance over different data aggregation levels (both spatial and temporal) should be investigated.

In the second problem, the emergence of ride-hailing services can help serve the growing transportation demand of our expanding cities, significantly improving the quality of city life and access to different places. From a spatio-temporal perspective, the study findings help us better understand human movement patterns. This study provides new insights on modeling travel behavior of ride-hailing service users. It shows a tremendous potential in predicting individual travel by using this emerging mode. Our results can also provide insights to develop high-fidelity simulations of on-demand service operations, which can further benefit develop services that



depend on ride-hailing. As a future application of these findings, we will focus on developing high-resolution generative models to forecast individual movements in cities.

## REFERENCES

- AGUILERA, A. & VOISIN, M. 2014. Urban form, commuting patterns and CO2 emissions: What differences between the municipality's residents and its jobs? *Transportation Research Part a-Policy and Practice*, 69, 243-251.
- ALAJALI, W., ZHOU, W., WEN, S. & WANG, Y. 2018. Intersection traffic prediction using decision tree models. *Symmetry*, 10, 386.
- ALESSANDRETTI, L., SAPIEZYNSKI, P., LEHMANN, S. & BARONCHELLI, A. 2017. Multi-scale spatio-temporal analysis of human mobility. *Plos One*, 12.
- ALONSO-MORA, J., SAMARANAYAKE, S., WALLAR, A., FRAZZOLI, E. & RUS, D. 2017a. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. *Proceedings of the National Academy of Sciences*, 114, 462-467.
- ALONSO-MORA, J., WALLAR, A. & RUS, D. Predictive routing for autonomous mobility-on-demand systems with ride-sharing. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017b. IEEE, 3583-3590.
- BAJARDI, P., POLETTO, C., RAMASCO, J. J., TIZZONI, M., COLIZZA, V. & VESPIGNANI, A. 2011. Human Mobility Networks, Travel Restrictions, and the Global Spread of 2009 H1N1 Pandemic. *Plos One*, 6.
- BANSAL, P. & KOCKELMAN, K. M. 2017. Forecasting Americans' long-term adoption of connected and autonomous vehicle technologies. *Transportation Research Part a-Policy and Practice*, 95, 49-63.
- BARIMANI, N., MOSHIRI, B. & TESHNEHLAB, M. 2012. State space modeling and short-term traffic speed prediction using kalman filter based on ANFIS. *International Journal of Engineering and Technology*, 4, 116.
- BROCKMANN, D., HUFNAGEL, L. & GEISEL, T. 2006. The scaling laws of human travel. *Nature*, 439, 462-465.

- CHEN, C., MA, J. T., SUSILO, Y., LIU, Y. & WANG, M. L. 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C-Emerging Technologies*, 68, 285-299.
- CHEN, D. Q., YAN, X. D., LIU, F., LIU, X. B., WANG, L. W. & ZHANG, J. C. 2019. Evaluating and Diagnosing Road Intersection Operation Performance Using Floating Car Data. *Sensors*, 19.
- CHEN, T. & GUESTRIN, C. XGBoost: reliable large-scale tree boosting system. *Proceedings of the 22nd SIGKDD Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2015. 13-17.
- CHEN, X. M., ZAHIRI, M. & ZHANG, S. 2017. Understanding ridesplitting behavior of on-demand ride services: An ensemble learning approach. *Transportation Research Part C: Emerging Technologies*, 76, 51-70.
- CONTRERAS, S. D. & PAZ, A. 2018. The effects of ride-hailing companies on the taxicab industry in Las Vegas, Nevada. *Transportation Research Part a-Policy and Practice*, 115, 63-70.
- DONG, Y. Q., WANG, S. F., LI, L. & ZHANG, Z. 2018. An empirical study on travel patterns of internet based ride-sharing. *Transportation Research Part C-Emerging Technologies*, 86, 1-22.
- DOWLING, R., SKABARDONIS, A., CARROLL, M. & WANG, Z. 2004. Methodology for measuring recurrent and nonrecurrent traffic congestion. *Transportation Research Record*, 1867, 60-68.
- EHMKE, J. F., MEISEL, S. & MATTFELD, D. C. 2012. Floating car based travel times for city logistics. *Transportation Research Part C-Emerging Technologies*, 21, 338-352.
- FRIEDMAN, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.
- GONZALEZ, M. C., HIDALGO, C. A. & BARABASI, A. L. 2008. Understanding individual human mobility patterns. *Nature*, 453, 779-782.

- HASAN, S., SCHNEIDER, C. M., UKKUSURI, S. V. & GONZALEZ, M. C. 2013a. Spatiotemporal Patterns of Urban Human Mobility. *Journal of Statistical Physics*, 151, 304-318.
- HASAN, S. & UKKUSURI, S. V. 2018. Reconstructing Activity Location Sequences From Incomplete Check-In Data: A Semi-Markov Continuous-Time Bayesian Network Model. *Ieee Transactions on Intelligent Transportation Systems*, 19, 687-698.
- HASAN, S., ZHAN, X. & UKKUSURI, S. V. Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*, 2013b. ACM, 6.
- HERRING, R., HOFLEITNER, A., ABBEEL, P. & BAYEN, A. Estimating arterial traffic conditions using sparse probe data. *13th International IEEE Conference on Intelligent Transportation Systems*, 2010. IEEE, 929-936.
- HORANONT, T., WITAYANGKURN, A., SEKIMOTO, Y. & SHIBASAKI, R. 2013. Large-Scale Auto-GPS Analysis for Discerning Behavior Change during Crisis. *Ieee Intelligent Systems*, 28, 26-34.
- HUANG, S.-H. & RAN, B. 2003. An application of neural network on traffic speed prediction under adverse weather condition. *University of Wisconsin--Madison*.
- HUANG, Z. R., LING, X. M., WANG, P., ZHANG, F., MAO, Y. P., LIN, T. & WANG, F. Y. 2018. Modeling real-time human mobility based on mobile phone and transportation data fusion. *Transportation Research Part C-Emerging Technologies*, 96, 251-269.
- JIA, Y., WU, J. & DU, Y. Traffic speed prediction using deep learning method. *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, 2016. IEEE, 1217-1222.
- JITTRAPIROM, P., CAIATI, V., FENERI, A. M., EBRAHIMIGHAREHBAGHI, S., ALONSO-GONZALEZ, M. J. & NARAYAN, J. 2017. Mobility as a Service: A Critical Review of Definitions, Assessments of Schemes, and Key Challenges. *Urban Planning*, 2, 13-25.

- JURDAK, R., ZHAO, K., LIU, J. J., ABOUJAOUDE, M., CAMERON, M. & NEWTH, D. 2015. Understanding Human Mobility from Twitter. *Plos One*, 10.
- KE, J., ZHENG, H., YANG, H. & CHEN, X. M. 2017. Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach. *Transportation Research Part C: Emerging Technologies*, 85, 591-608.
- LAVIERI, P. S. & BHAT, C. R. 2019. Modeling individuals' willingness to share trips with strangers in an autonomous vehicle future. *Transportation Research Part a-Policy and Practice*, 124, 242-261.
- LENG, B., DU, H., WANG, J. Y., LI, L. & XIONG, Z. 2016. Analysis of Taxi Drivers' Behaviors Within a Battle Between Two Taxi Apps. *Ieee Transactions on Intelligent Transportation Systems*, 17, 296-300.
- LU, X., BENGTSSON, L. & HOLME, P. 2012. Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences of the United States of America*, 109, 11576-11581.
- MA, X., DAI, Z., HE, Z., MA, J., WANG, Y. & WANG, Y. 2017a. Learning traffic as images: a deep convolutional neural network for large-scale transportation network speed prediction. *Sensors*, 17, 818.
- MA, X., TAO, Z., WANG, Y., YU, H. & WANG, Y. 2015. Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies*, 54, 187-197.
- MA, X. L., LIU, C. C., WEN, H. M., WANG, Y. P. & WU, Y. J. 2017b. Understanding commuting patterns using transit smart card data. *Journal of Transport Geography*, 58, 135-145.
- MYUNG, J., KIM, D.-K., KHO, S.-Y. & PARK, C.-H. 2011. Travel time prediction using k nearest neighbor method with combined data from vehicle detector system and automatic toll collection system. *Transportation Research Record*, 2256, 51-59.

- PAN, G., QI, G. D., ZHANG, W. S., LI, S. J., WU, Z. H. & YANG, L. T. 2013. Trace Analysis and Mining for Smart Cities: Issues, Methods, and Applications. *Ieee Communications Magazine*, 51, 120-126.
- PENG, C. B., JIN, X. G., WONG, K. C., SHI, M. X. & LIO, P. 2012. Collective Human Mobility Pattern from Taxi Trips in Urban Area. *Plos One*, 7.
- QIAN, X. & UKKUSURI, S. V. 2015. Spatial variation of the urban taxi ridership using GPS data. *Applied Geography*, 59, 31-42.
- RAHMAN, R. & HASAN, S. Short-Term Traffic Speed Prediction for Freeways During Hurricane Evacuation: A Deep Learning Approach. 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 2018. IEEE, 1291-1296.
- RASHIDI, T. H., ABBASI, A., MAGHREBI, M., HASAN, S. & WALLER, T. S. 2017. Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges. *Transportation Research Part C-Emerging Technologies*, 75, 197-211.
- SUN, Y. R., FAN, H. C., LI, M. & ZIPF, A. 2016. Identifying the city center using human travel flows generated from location-based social networking data. *Environment and Planning B-Planning & Design*, 43, 480-498.
- SUZUKI, T. & LEE, S. 2012. Jobs-housing imbalance, spatial correlation, and excess commuting. *Transportation Research Part a-Policy and Practice*, 46, 322-336.
- TIAN, G. J., WU, J. G. & YANG, Z. F. 2010. Spatial pattern of urban functions in the Beijing metropolitan region. *Habitat International*, 34, 249-255.
- VAN HINSBERGEN, C., HEGYI, A., VAN LINT, J. & VAN ZUYLEN, H. 2011. Bayesian neural networks for the prediction of stochastic travel times in urban networks. *IET intelligent transport systems*, 5, 259-265.
- VANICHRUJEE, U., HORANONT, T., PATTARA-ATIKOM, W., THEERAMUNKONG, T. & SHINOZAKI, T. Taxi Demand Prediction using Ensemble Model Based on RNNs and XGBOOST. 2018 International Conference on Embedded Systems and Intelligent

Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES), 2018. IEEE, 1-6.

VELOSO, M., PHITHAKKITNUKON, S., BENTO, C., FONSECA, N. & OLIVIER, P. Exploratory study of urban flow using taxi traces. First Workshop on Pervasive Urban Applications (PURBA) in conjunction with Pervasive Computing, San Francisco, California, USA, 2011.

WANG, H., LIU, L., QIAN, Z., WEI, H. & DONG, S. 2014. Empirical mode decomposition–autoregressive integrated moving average: hybrid short-term traffic speed prediction model. *Transportation Research Record*, 2460, 66-76.

WANG, W. J., PAN, L., YUAN, N., ZHANG, S. & LIU, D. 2015. A comparative analysis of intra-city human mobility by taxi. *Physica a-Statistical Mechanics and Its Applications*, 420, 134-147.

WEN, H., SUN, J. & ZHANG, X. 2014. Study on traffic congestion patterns of large city in China taking Beijing as an example. *Procedia-Social and Behavioral Sciences*, 138, 482-91.

WEN, J., NASSIR, N. & ZHAO, J. H. 2019. Value of demand information in autonomous mobility-on-demand systems. *Transportation Research Part a-Policy and Practice*, 121, 346-359.

WESOLOWSKI, A., EAGLE, N., TATEM, A. J., SMITH, D. L., NOOR, A. M., SNOW, R. W. & BUCKEE, C. O. 2012. Quantifying the Impact of Human Mobility on Malaria. *Science*, 338, 267-270.

WILLIAMS, B. M. & HOEL, L. A. 2003. Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results. *Journal of transportation engineering*, 129, 664-672.

WOLF, J., GUENSLER, R. & BACHMAN, W. 2001. Elimination of the travel diary - Experiment to derive trip purpose from global positioning system travel data. *Transportation Data and Information Technology*, 125-134.

- WU, C.-H., HO, J.-M. & LEE, D.-T. 2004. Travel-time prediction with support vector regression. *IEEE transactions on intelligent transportation systems*, 5, 276-281.
- WU, J., JIANG, C., HOUSTON, D., BAKER, D. & DELFINO, R. 2011. Automated time activity classification based on global positioning system (GPS) tracking data. *Environmental Health*, 10, 101.
- XIA, D., WANG, B., LI, H., LI, Y. & ZHANG, Z. 2016. A distributed spatial-temporal weighted model on MapReduce for short-term traffic flow forecasting. *Neurocomputing*, 179, 246-263.
- YAO, B., CHEN, C., CAO, Q., JIN, L., ZHANG, M., ZHU, H. & YU, B. 2017. Short-term traffic speed prediction for an urban corridor. *Computer-Aided Civil and Infrastructure Engineering*, 32, 154-169.
- ZHAN, X., HASAN, S., UKKUSURI, S. V. & KAMGA, C. 2013. Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C: Emerging Technologies*, 33, 37-49.
- ZHANG, F., ZHU, X., HU, T., GUO, W., CHEN, C. & LIU, L. 2016. Urban link travel time prediction based on a gradient boosting method considering spatiotemporal correlations. *ISPRS International Journal of Geo-Information*, 5, 201.
- ZHANG, J. P., WANG, F. Y., WANG, K. F., LIN, W. H., XU, X. & CHEN, C. 2011. Data-Driven Intelligent Transportation Systems: A Survey. *Ieee Transactions on Intelligent Transportation Systems*, 12, 1624-1639.
- ZHANG, X. H., XU, Y., TU, W. & RATTI, C. 2018. Do different datasets tell the same story about urban mobility - A comparative study of public transit and taxi usage. *Journal of Transport Geography*, 70, 78-90.
- ZHAO, K., MUSOLESI, M., HUI, P., RAO, W. X. & TARKOMA, S. 2015. Explaining the power-law distribution of human mobility through transportation modality decomposition. *Scientific Reports*, 5.



- ZHAO, L., ZHOU, Y., LU, H. & FUJITA, H. 2019. Parallel computing method of deep belief networks and its application to traffic flow prediction. *Knowledge-Based Systems*, 163, 972-987.
- ZHAO, Z., KOUTSOPOULOS, H. N. & ZHAO, J. H. 2018. Individual mobility prediction using transit smart card data. *Transportation Research Part C-Emerging Technologies*, 89, 19-34.
- ZHENG, L. J., XIA, D., ZHAO, X., TAN, L. Y., LI, H., CHEN, L. & LIU, W. N. 2018. Spatial-temporal travel pattern mining using massive taxi trajectory data. *Physica a-Statistical Mechanics and Its Applications*, 501, 24-41.
- ZHOU, Z.-H. 2012. *Ensemble methods: foundations and algorithms*, Chapman and Hall/CRC.