2023

Modeling Individual Activity and Mobility Behavior and Assessing Ridesharing Impacts Using Emerging Data Sources

Jiechao Zhang
University of Central Florida

MODELING INDIVIDUAL ACTIVITY AND MOBILITY BEHAVIOR AND ASSESSING RIDESHARING IMPACTS USING EMERGING DATA SOURCES

by

JIECHAO ZHANG

B.Sc. Beijing Jiaotong University, 2015

M.Sc. University of Central Florida, 2020

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
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

2023

Major Professor: Samiul Hasan

ABSTRACT

Predicting individual mobility behavior is one of the major steps of transportation planning models. Accurate prediction of individual mobility behavior will be beneficial for transportation planning. Although previous studies have used different data sources to model individual mobility behaviors, they have several limitations such as the lack of complete mobility sequences and travel mode information, limiting our ability to accurately predict individual movements. In recent years, the emergence of GPS-based floating car data (FCD) and on-demand ride-hailing service platforms can provide innovative data sources to understand and model individual mobility behavior. Compared to the previously used data sources such as mobile phone and social media data, mobility data extracted of the new data sources contain more specific, detailed, and longitudinal information of individual travel mode and coordinates of the visited locations. This dissertation explores the potential of using GPS-based FCD and on-demand ride-hailing service data with different modeling techniques towards understanding and predicting individual mobility and activity behaviors and assessing the ridesharing impacts through three studies.

Keywords: *Individual Activity Behavior; Individual Mobility Prediction; Intelligent Transportation System; IOHMM; Big Data*

EXTENDED ABSTRACT

First, we developed a method to infer individual activity participation type and developed an individual-level activity prediction model—an input-output hidden Markov model (IOHMM)—to generate the activity sequences, durations, and destinations of consecutive trips made by an

individual over a day. Two datasets including vehicle trajectory and point of interest (POI) data were used in this study. The developed IOHMM model can generate the activity sequence with about 90% accuracy for afternoon, evening, and night and more than 80% accuracy for morning and midday periods. With the generated activity type, the model can also predict activity type and locations accurately. The model presents strong interpretability since we can obtain the activity probabilities from model parameters, such as initial, transition and emission parameters.

Second, using large-scale data extracted from 50,000 ride-hailing service users' anonymized mobility records, we applied a multi-layer hidden Markov model to predict on-demand ride-haling service trips for each individual considering the heterogeneity of the travel purposes, including the trip decision, the number of daily trips, the origin, and the destination. The results show that the trip decision model can achieve up to 65% accuracy. The origin and destination prediction model can work well on the commute-based users which can achieve nearly 70% accuracy. To better evaluate the performance of origin prediction model, we discovered that the accuracy of origin prediction model is strongly correlated with the predictability of a mobility sequence.

Third, an agent-based model was developed to simulate the influence of a ridesharing strategy with real-world data available from a ride-hailing service. To better satisfy the passengers' requests, we proposed a real-time vehicle-passenger matching algorithm both for the ridesharing and non-ridesharing scenarios. The results show that the ridesharing system can decrease up to 55% of the fleet size and 51% of the unoccupancy rate compared to a system without any ridesharing option available. The average waiting time of the ridesharing system can be lowered by 79.2% with a fleet size of 300. Besides, the ridesharing system can decrease between 7.6% and 65.6% of the total vehicle kilometers traveled and between 7.2% and 40% of carbon dioxide

(CO_2) emissions considering different fleet sizes. In addition, the results indicate that the ridesharing strategy can work better in high demand areas and peak hours.

ACKNOWLEDGMENTS

I would like to convey my heartiest gratitude to my honorable supervisor Dr. Samiul Hasan for his continuous support and supervision throughout my Ph.D. journey. His guidance, resourceful insights, and wisdom directed me on the way to complete my dissertation in time. I would also like to acknowledge the support and encouragement from my family, friends, and all the people I look for whenever I go through difficult times.

TABLE OF CONTENT

LIST OF FIGURES x
LIST OF TABLES xii
CHAPTER 1: INTRODUCTION 1
1.1 Background 1
1.2 Motivation 3
1.2.1 Individual Activity Behavior Prediction with Vehicle Trajectory Data 3
1.2.2 Intersections Individual Mobility Behavior Prediction for Ride-hailing Service Users 4
1.2.3 Ridesharing Effects 5
1.3 Dissertation Objectives 6
1.4 Contributions 7
1.5 Structure of the Dissertation 8
CHAPTER 2: MODELING INDIVIDUAL ACTIVITY BEHAVIOR USING VEHICLE TRAJECTORY DATA 10
2.1 Introduction 10
2.2 Literature Review 12
2.3 Study Area and Data Description 15
2.4 Input-Output Hidden Markov Model (IOHMM) 21
2.5 Results 25
2.5.1 Inferring activity type 25
2.5.2 IOHMM model results 28
2.6 Conclusions 39

CHAPTER 3: PREDICTING INDIVIUDAL MOBILITY BEHAVIOR OF RIDE-HAILING SERVICE USERS CONSIDERING HETEROGENEITY OF TRIP PURPOSES 42
3.1 Introduction 42
3.2 Literature Review 44
3.3 Data and Methods 46
3.3.1 Data Description 46
3.3.2 Data Exploration 48
3.3.3 Clustering Users 50
3.3.4 Multi-Layer Hidden Markov Model 53
3.3.5 Predictability of Mobility 59
3.4 Empirical Results 60
3.4.1 Mobility Patterns 60
3.4.2 Trip Decision Prediction 62
3.4.3 Next Origin and Destination Prediction 65
3.4.4 Temporal Patterns of Model Accuracy 68
3.4.5 Predictability vs. Model Accuracy 70
3.5 Conclusions 71
CHAPTER 4: ASSESSING THE IMPACTS OF A REAL-TIME RIDESHAING SYSTEM USING AN AGENT-BASED SIMULATION MODEL 74
4.1 Introduction 74
4.2 Literature Review 75
4.3 Data and Methods 79
4.3.1 Agent-based Model 79
4.3.2 Matching Algorithm - Base Scenario 81
4.3.3 Matching Algorithm - Ridesharing Scenario 81

4.3.4 Empty Service Vehicle Relocation 86
4.4 Simulation Results 87
4.4.1 Fleet Size vs. Maximum Waiting Time 88
4.4.2 Fixed Fleet Size 91
4.4.3 Vehicle Kilometers Traveled (VKT) 93
4.4.4 Environment Benefits 94
4.5 Conclusions 95
CHAPTER 5: CONCLUSIONS 98
5.1 Summary of Major Results 99
5.2 Limitations and Future Research Directions 101
REFERENCES 104

LIST OF FIGURES

Figure 2.1 Sample trips from the telematic data collection survey 16

Figure 2.2 The visited coordinates (a) vs. visited place or point of interest (b) 17

Figure 2.3 (a) The distribution of visited places number; (b) The probability of rank of visited places 18

Figure 2.4 Grid division of the study area 21

Figure 2.5 The framework of input-output hidden Markov model (IOHMM) 22

Figure 2.6 Distribution of start time for different activities 26

Figure 2.7 Activity duration distribution for each activity type 27

Figure 2.8 Individual daily activity sequence with activity types for two vehicles (a) vehicle 9 and (b) vehicle 10 28

Figure 2.9 Results of number of generated and actual activities by type 31

Figure 2.10 Results of number of generative activity duration vs. real activity types 32

Figure 2.11 Distribution of model accuracy for activity location prediction 33

Figure 2.12 Distribution of initial probability in different time periods 35

Figure 2.13 Distribution of transition probability in different time periods 38

Figure 2.14 Ground truth of activity sequence vs. generated activity sequence: (a) ground truth of vehicle 9 activity sequence and (b) generated vehicle 9 activity sequence 39

Figure 3.1 The study region: area inside Beijing 6th ring road 48

Figure 3.2 Exploratory data analysis: (a) the distribution of the number of trips; (b) the distribution of the number of active days; (c) the distribution of the number of visited places; (d) the distribution of rank of visited places in log-log scale 50

Figure 3.3 The distribution of rank of visited places in different hour: (a) home-based users; (b) work-based users; (c) commute-based users; (d) random users. 53

Figure 3.4 Multi-layer hidden Markov model framework..... 55

Figure 3.5 Spatio-Temporal patterns of ride-hailing service: (a) the distribution of jump length; (b) the distribution of radius of gyration; (c) the distribution of hourly trip generation 62

Figure 3.6 Trip decision prediction model results: (a) home-based users; (b) work-based users; (c) commute-based users; (d) random users 64

Figure 3.7 Origin prediction model results: (a) home-based users; (b) work-based users; (c) commute-based users; (d) random users 66

Figure 3.8 Destination prediction model results: (a) home-based users; (b) work-based users; (c) commute-based users; (d) random users 67

Figure 3.9 Model accuracy for different temporal features for each cluster 70

Figure 3.10 The results of mobility behavior predictability 71

Figure 4.1 Study area..... 80

Figure 4.2 Ridesharing simulation modeling framework 82

Figure 4.3 Potential routes for ridesharing strategy..... 84

Figure 4.4 Empty service vehicle relocation strategy 87

Figure 4.5 Daily (A) and Hourly (B) distribution of trip number..... 88

Figure 4.6 Average waiting time (A) and average travel time (B) with different fleet size.. 92

Figure 4.7 VMT with different fleet size..... 94

LIST OF TABLES

Table 2.1 Activity type identification rules 20

Table 2.2 Results of comparison of activity number 29

Table 3.1 Detailed Data Attributes 47

Table 3.2 Emission parameters of daily trip number of trip decision prediction model 65

Table 4.1 Relationship between the maximum waiting time and fleet size 90

Table 4.2 Relationship between the maximum waiting time and unoccupancy rate 90

Table 4.3 Ridesharing rate for different fleet size 93

Table 4.4 CO_2 emissions for different fleet sizes 95

CHAPTER 1: INTRODUCTION

1.1 Background

Analyzing spatio-temporal patterns of individual mobility is one of the core elements of transportation analysis and modeling. Thus, understanding and modeling individual mobility behavior, from different perspectives and data sources, is beneficial to many transportation applications such as transportation planning [1], intelligent transportation system (ITS) [2] design, development of smart cities [3], and traffic management [4]. Traditionally, individual mobility behavior analysis is mainly based on travel surveys [5]. Although travel survey methods have evolved from traditional pen-and-paper based data collection to nowadays web and smartphone-based data collection, it also has disadvantages due to low-efficiency and high-cost limiting the ability to understand and model the individual mobility behaviors [6].

In recent years, to understand individual mobility behavior, various large-scale high-resolution datasets with varying capabilities have been used [7, 8]. For instance, call details records (CDR) extracted from mobile phones can provide information of human movement behavior in different spatio-temporal scales [9-12]. Besides, massive social media data [13] extracted from online social media platforms are also used in the individual mobility and activity analysis. However, both CDR and social media data have their own limitations in understanding human mobility behavior. First, the data can be recorded only when the user is utilizing the mobile device or log in the social media platform which will be challenging to provide the whole mobility or activity sequence of an individual. Second, we cannot identify the travel mode when an individual makes a trip or have an activity which limits the application of the data for solving transportation problems. Third, both the CDR and social media data are difficult to provide the precise location visited by individuals. For CDR data, since the data is collected based on tower

locations, we can only obtain the approximate area of the visited location. For social media data, since not all the users will share their geolocation information, it is difficult to get enough location-based data. It will provide more valuable insights if we can learn the patterns of human activity and mobility patterns with the data sources that contain more precise transportation related information such travel mode and specific origin and destination of a trip.

With the rapid development of GPS-based trajectory data and on-demand ride-hailing service platform, massive innovative transportation related data sources have been available for understanding and modeling mobility behavior [14]. Both GPS-based trajectory data and on-demand ride-hailing data can overcome some of the limitations of previous data sources which can be beneficial in revealing the individual mobility patterns. For instance, GPS-based trajectory data can provide information of travel mode, the timestamp of the activities or movements, the complete mobility sequence and the precise locations of the visited places. The on-demand ride-hailing service data can also share the information of time, mode, and specific visited location.

In addition, on-demand ride-hailing platforms like Uber and Lyft offer convenient, flexible transportation options for users, allowing them to book a ride anytime, anywhere through a simple smartphone app. This innovative ride-hailing service has created opportunities for individuals to use ridesharing as a means of transportation, since passengers with similar routes and schedules can share a ride. Ride-sharing services can help reduce traffic congestion, decrease air pollution, and reduce the overall carbon footprint of transportation systems [15, 16]. Additionally, by reducing the number of vehicles on the road, ridesharing can help to save valuable resources, such as fuel and minerals [17], which are used to produce and maintain

vehicles. Overall, the trend towards ridesharing is a positive development for both the environment and society.

This dissertation utilizes emerging mobility data such GPS-based trajectory and on-demand ride-hailing service data for developing new methodologies to understand and model individual mobility behavior. Furthermore, using on-demand ride-hailing service data, it develops an agent-based simulation model for real-time ridesharing strategies and evaluates the impacts of implementing such strategies at a system level. The understanding of human mobility behavior and the impacts of ridesharing strategies will enable us to make better transportation planning and traffic management tools and better transportation policies, which will increase the efficiency of transportation networks.

1.2 Motivation

The overarching goal of this dissertation is to better understand human mobility behavior and the effects of ridesharing strategy using GPS-based FCD and on-demand ride-hailing service data. To do that, we work on three inter-dependent studies. The motivation of choosing these studies is discussed below.

1.2.1 Individual Activity Behavior Prediction from Vehicle Trajectory Data

Predicting the next activity and the corresponding location of individual mobility behavior is one of the important topics in human movement or travel behavior modeling. Although previous researchers have utilized data from mobile phone records [18-24], smart card transactions [25, 26] and social media posts [27, 28] to understand mobility patterns and develop individual mobility prediction model, the use of other data sources is also gaining interest. Research are also interested to use the data sources that can provide more transportation related information,

since both mobile phone or social media data have limits in observing precise mobility behavior, and smart card data cannot provide continuous human trajectories.

Since most of the traffic are generated by private vehicle trips, which is the primary mode of transportation for the people in the USA [29], a generative model for individual activity behaviors using trajectory information of individual vehicles (i.e., private car) will be essential for future urban and transportation planning and management. Although a variety of studies have been conducted for modeling individual activity behaviors, there exist several research gaps. First, due to a lack of long-term individual-level vehicle trajectories, few studies have focused on identifying the activities from individual trajectories. Second, no previous studies have developed a generative model for individual activity sequence with the trajectories available from private vehicles.

The above limitations motivate us to develop a generative model for individual activity sequence using GPS-based individual trajectory data. A generative model for individual activity sequence using GPS-based individual trajectory data can provide a powerful tool for understanding human behavior and generating personalized recommendations.

1.2.2 Individual Mobility Behavior Prediction from Ride-hailing Service Data

Previous studies [30-35] developed algorithms for the next activity prediction or next location prediction. However, for some specific travel modes such as on-demand ride-hailing service trips, there is no existing model for predicting trip decisions and generating trip sequences at an individual level. The trip decision of ride-hailing service trips represents whether or not an individual will use the ride-hailing service trip given a specific contextual information such as the day of the week and weather; this element is also important to understand individual mobility behavior by a specific travel mode. By modeling individual trip decisions, a transportation

network company can design appropriate strategies and offer incentives to manage the demand for ride-hailing services and improve rider experience.

Individual mobility behaviors show both regularities and uncertainties. For instance, daily commuting behaviors of individuals show strong regularities. However, an individual may also explore new places indicating the uncertainty in mobility behavior. Several previous studies have investigated the predictability of human mobility using mobile phone data. However, most of the previous studies focused on predictability when forecasting the individual location in the next time step. Such an approach cannot reflect the predictability of individual mobility in a realistic way since individuals may stay in the same place for a longer period of time.

The above limitations have motivated us to investigate alternative data sources and approaches for understanding human mobility behavior. In recent years, the rapid development of on-demand ride-hailing service platforms provide us more innovative transportation related data which created an opportunity towards developing mobility prediction models and analyze the predictability of mobility sequence of trips made by private vehicles and on-demand ride-hailing services.

1.2.3 Impacts of Ridesharing

In recent years, the wide adoption of on-demand ride-hailing service platforms such as Uber, Lyft and Didi has significantly influenced individual mobility options in their daily life. The development of smart-phone based technology and the on-demand ride-haling service platforms have enabled ridesharing services which will not only reduce traffic volume in the transportation network but also have significant environmental benefits [17]. Besides, while a ride-sharing strategy can reduce the number of vehicles on the road, it can also have negative effects. One of the main drawbacks is that the detour distance can increase as the ride-sharing vehicle may need

to pick up or drop off multiple passengers along the way. This can lead to longer travel times and decreased efficiency for the passengers. Additionally, the detour distance can also result in increased fuel consumption and higher emissions, which can have a negative impact on the environment.

Moreover, in some cases, ridesharing can lead to increased traffic congestion, especially during peak hours, as multiple vehicles are picking up and dropping off passengers along the same routes. To minimize these negative effects, it's important to continuously optimize their algorithms and routing strategies to minimize detour distances and reduce the number of vehicles on the road. In order to fully understand the potential impacts of ridesharing, it is essential to evaluate the ridesharing strategy using real-world data. This can help to accurately measure the benefits and drawbacks of ridesharing and inform future decisions about transportation policy and planning.

1.3 Dissertation Objectives

This dissertation presents studies to improve our understanding of individual mobility behavior and to fill the gaps in existing studies using GPS-based floating car data and on-demand ride-hailing service data. Besides, we assess the impacts of a proposed ridesharing strategy at the level of a system. The dissertation has the following specific objectives:

i. Develop a method to identify the activity type of individual mobility sequence. Then, develop an individual-level mobility prediction model, to predict the trajectories of daily consecutive vehicle trips.
ii. Develop a model to predict the ride-hailing service trips including the trip decision, the daily trip number, and the origin and destination of a trip; analyze mobility patterns with

on-demand ride-hailing service data; and determine the predictability of individual mobility sequence.

iii. Develop a method to evaluate the traffic and environmental impacts of a real-time ridesharing strategy on urban traffic network using ride-hailing service data. To do that, we need to build a real-time ridesharing system using an agent-based simulation model to extract the trajectories of passengers and vehicles which can provide valuable information to evaluate the potential environment and transportation impacts.

1.4 Contributions

This dissertation has made the following contributions:

i. This dissertation augments vehicle trajectory data with a database of point of interests (POI) to infer activity type from the POI category and the start time and duration of an activity. Besides, it develops an input-output hidden Markov model (IOHMM) using real-world vehicle trajectory data to predict the type and location of the next activity and generate activity sequence. To the best of our knowledge, this is the first study to develop a generative model of activity sequence from vehicle trajectory data.

ii. In addition, it develops an individual-level mobility prediction model, to predict the trip making behavior of ride-hailing service users. To the best of our knowledge, this is one of the few studies predicting individual mobility behavior using massive ride-hailing service data which will not only provide insights for urban planning and traffic management applications but also can improve the services offered by transportation network companies.

iii. This dissertation investigates the entropy and predictability of the mobility sequence of ride-hailing service users and validates the performance of individual mobility prediction

model using the concept of predictability. The method of predictability can be applied to improve the performance of models predicting individual mobility behavior.

iv. This dissertation develops an agent-based simulation model that can evaluate the traffic and environmental impacts of ridesharing strategies on urban traffic networks using real-world ride-hailing service data.

1.5 Structure of the Dissertation

Chapter 2 presents the study that applied an input-output hidden Markov model (IOHMM) to develop a generative individual activity behavior model using the data extracted from a telematic vehicle survey conducted in Ann Arbor, Michigan. We identified the activity types based on the spatio-temporal scales of human mobility and tested the proposed model using the individual mobility sequence generated from the continuous vehicle trips.

Chapter 3 presents an innovative multi-layer Markov chain-based model to predict individual mobility behaviors using the data extracted from Didi platform which is one of the largest on-demand ride-hailing platforms of the world. In this research, we focus on predicting the next ride-hailing service trips, including the trip decision, the daily trip number, the origin and destination of the trip. Besides, we use the predictability concept calculated by entropy to evaluate the performance of the mobility sequence prediction model.

Chapter 4 presents an agent-based simulation model to evaluate the travel and environmental impacts of ride-sharing strategies. Using a real-time agent-based simulation model and incorporating real-world data from a ride-hailing service platform and mapping API, we aim to provide valuable insights into the benefits and drawbacks of ridesharing. The use of real demand data from a ride-hailing service platform and traffic condition data from Google

Map API will help to increase the realism of the simulation and make the results more representative of the real-world situation.

Finally, Chapter 5 concludes the dissertation by discussing the overall findings of the chapters, stating the limitations of the studies, and providing directions for future research.

CHAPTER 2: MODELING INDIVIDUAL ACTIVITY BEHAVIOR USING VEHICLE TRAJECTORY DATA

2.1 Introduction

Human activity is one of the most fundamental drivers of transportation demand [21]. Previous research has identified strong regularities in individual movement patterns [36, 37]. A better understanding of human activity participation behavior will facilitate improvement in various sectors including traffic operations and management [38], implementation of mobility-as-a-service (MaaS), ridesharing [39-41], and urban transportation planning [1]. In recent years, data-driven methods have been widely applied to analyze and model human mobility patterns [42]. In the pioneering study, using mobile phone data, Gonzalez et al. [23] revealed that individual mobility behavior shows strong regularities in a spatio-temporal scale. Additionally, data from social media [43, 44], public transportation [25, 45], taxi GPS [46, 47], and smartphones [42] have also been utilized to understand human mobility behavior.

Predicting individual mobility and activity behavior such as next activity location [37] or next trip [25] is an important topic in travel behavior modeling. The data used for individual mobility prediction can be broadly divided into two groups – *extrinsic mobility data* and *intrinsic mobility data* [25]. Data from smartphones and social media are defined as *extrinsic mobility data* since these data are passively generated as a consequence of activity participation process. Therefore, information about travel modes and visited locations are not directly available from extrinsic mobility data. Additionally extrinsic mobility data cannot capture the continuous trace of human mobility sequence. Mobility information from extrinsic data is recorded only when an individual uses a smartphone or an application (also referred to as app such as Facebook or Twitter), resulting in non-consecutive mobility sequence. As such, using extrinsic mobility data

without any correction can lead to biased model parameters and inaccurate prediction of activity attributes such as duration and frequency.

On the other hand, *intrinsic mobility data* are directly extracted from the mobility behavior associated with specific travel modes, such as public transport, taxi, or bike sharing systems. This data can provide precise information on individual movement including visited locations, travel durations, and activity start and end times. Intrinsic mobility data are frequently used in transportation system modeling. However, most intrinsic data rely on traditional survey methods that require direct input from the respondents. Since individuals need to actively participate in the data collection process, such data cannot capture mobility and activity patterns beyond a couple of days, with most surveys collecting mobility behavior for a weekday [48]. However, previous research has identified that individual mobility information from a single day cannot capture the true trends in mobility patterns necessary for shaping transportation system performance [49]. In recent years, the rapid development of connected vehicles has shown the great potential for collecting massive individual mobility data through continuous logging of successive vehicle trips. Such datasets have significant potential for understanding human movement patterns since they preserve the continuity of visited locations (unlike smartphone data) and span across multiple days (unlike traditional survey data).

In this paper, we use vehicle trajectory data collected from Ann Arbor, Michigan to develop a predictive model of individual activity behavior. We first investigate the spatio-temporal patterns of vehicle users' mobility behavior to understand their travel distance and frequency. Based on the historical sequence of an individual, we identify activity types using point of interest (POI) information, activity start time, and activity duration. To predict individual activity and mobility behaviors, we develop an input-output hidden Markov model

(IOHMM). In the IOHMM model, activity types are considered as the hidden states, the inputs contain contextual information, such as time of the day and day of the week, and the output contains information on activity location.

The contributions of this research can be summarized as follows:

- The study augments vehicle trajectory data with a database of point of interests (POIs) to infer the activity type from a POI category and the start time and duration of an activity.
- The study develops an input-output hidden Markov model (IOHMM) using real world vehicle trajectory data to predict the type and location of the next activity and generate activity sequence using individual-level vehicle trajectory data. To the best of our knowledge, this is the first study to develop a generative model of activity sequence from vehicle trajectory data.
- The study uses real-world vehicle trajectory data collected from a city to train the IOHMM model at an individual level, evaluates model performance with respect to model accuracy and precision, and provides insights on how the results from such a model can be interpreted.

2.2 Literature Review

In recent years, many studies have been conducted to understand individual mobility behaviors [50]. Existing individual mobility research can be grouped into three categories: analyzing individual mobility patterns [23, 43, 45, 51-54], predicting individual mobility behaviors [25, 31, 35, 37, 38, 55, 56], and revealing uncertainty and predictability of human movement [21, 57-62].

The analysis of individual mobility patterns has received extensive attention since massive human movement data have become available. The widespread adoption of mobile devices and emerging technologies has generated high-quality mobility data from smartphone usage [14], social media platforms [43], smart card transactions [45], and GPS observations [63],

providing insights into human mobility patterns. For instance, analyzing mobile phone data, Gonzalez et al. [63] found that human movement has strong regularities with a higher probability associated with more frequently visited places. Zhao et al. discovered that the power law distribution of individual travel distance can be decomposed into multiple lognormal distributions specific to a travel mode. Similarly, using subway smart card transaction data, Hasan et al. [45] found that the distribution of the rank of visited locations follows Zipf's law when the rank is high. Using mobile phone data, Kang et al. [64] showed that the intra-urban travel generation of individuals follows an exponential distribution. These studies mainly focused on understanding human mobility patterns by fitting movement data through appropriate statistical distributions.

Since individual movement patterns have both regularities and uncertainties, researchers have analyzed the ability to forecast the future movement of individuals [21]. In previous studies, researchers mainly used entropy to capture the value of the predictability of individual mobility. Song et al. [21] analyzed 50,000 mobile phone users' mobility patterns and found a 93% predictability for individual mobility behaviors. Lu et al. [58] also used mobile phone call detail records (CDR) data to estimate the predictability of individual mobility behavior. Based on the entropy value, it was shown that the theoretical maximum predictability of human mobility can reach 88%. However, these two studies only focused on the predictability of visiting a location in the next time step, which leads to a high value of predictability as individuals tend to stay longer at a visited location. In the study [62], using 604 individual GPS traces data, the paper analyzed the predictability of forecasting the next visited location. The study reported 71% predictability in forecasting the next visited location which is substantially less than the previous study. It means that the next location is more difficult to predict than the location in the next time step.

Although the concept of predictability can inform us how much human mobility can be forecast, appropriate data-mining algorithms can learn individual mobility behaviors and predict the actual individual movement.

Many predictive models have been developed using emerging mobility data [58]. Such models include neural networks [65, 66], generic algorithms-based models [35], Bayesian *n*-gram models [25], and Markov based models [31, 37, 58, 67-70]. These studies also utilized multiple types of datasets including mobile phone data [58], smart card data [25], simulation data [68], and GPS data [67]. Although Markov chain models have achieved a great performance in predicting individual mobility, these models cannot capture the influence of mobility contexts such as time of the day and day of the week. However, such temporal information usually ignored in the Markov chain model is likely to have a significant influence on individual activity behavior. Lv et al. [31] developed a Hidden Markov Model (HMM) for predicting the point of interest from historical trajectory data with an accuracy ranging between 20% and 70% depending on the travel behavior of an individual. The individuals who have regular life patterns have the highest prediction accuracy (around 70% for all periods). For the other people who always move (day postman, party person, or hard postman), the prediction accuracy was less than 50% from 7:00 am to 8:00 pm. Zong et al. [67] combined multinomial logit and Markov chain model to predict the destination of only 10 individuals using multi-day GPS data. The model has 90% accuracy for weekday trips and 85% accuracy for weekends.

Although a variety of methods have been proposed for modeling individual activity behaviors, there are still several research gaps. First, due to a lack of long-term individual-level vehicle trajectories, very few studies have focused on identifying the activities from individual trajectories. Second, no previous studies have developed a generative model for individual

activity sequence using individual vehicles (private car) trajectory information, which is the primary mode of transportation for people in the USA [29]. In this paper, we develop an individual-level model to generate individual activity behavior including activity type and activity location using continuous vehicle trajectory data.

2.3 Study Area and Data Description

This study has used the data gathered from a telematic data collection effort conducted by the University of Michigan, Transportation Research Institute and Argonne National Laboratory between May 2017 and November 2018. The survey recruited around 500 vehicles and the data were collected over the 18 months recording almost 8M vehicle miles. The information collected during the survey could be grouped into three categories: GPS information, fuel information, battery information, and the time stamp of the GPS information. The GPS information includes latitude, longitude, altitude of the vehicle, and the number of satellites used. This study builds on the GPS information collected during the survey. The initial survey logged the GPS information every 5 seconds which was reduced to every 3 seconds by the end of the year 2017. A map-matching algorithm was applied to the raw data to identify the routes and the trip start/end points. To anonymize the data, a stochastic algorithm was applied to randomize the trip start and end location within an admissible radius. All the data were stored in a server owned by Argonne National Laboratory and the models presented in this paper were run in that server. Individual mobility data was accessible to Argonne Lab researchers only. **Figure 2.1** shows some of the sample trip trajectories collected in May 2018 with the red and blue dots indicating the trip start and end points, respectively.

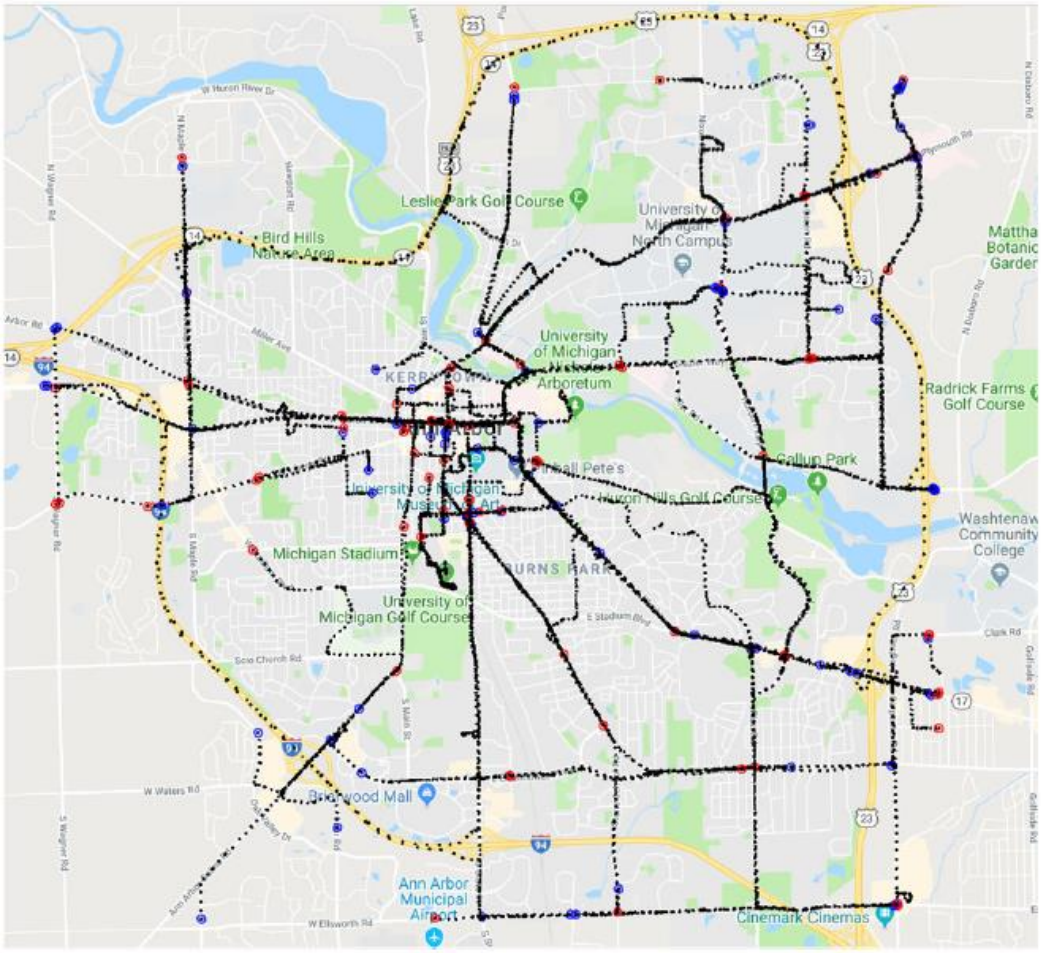
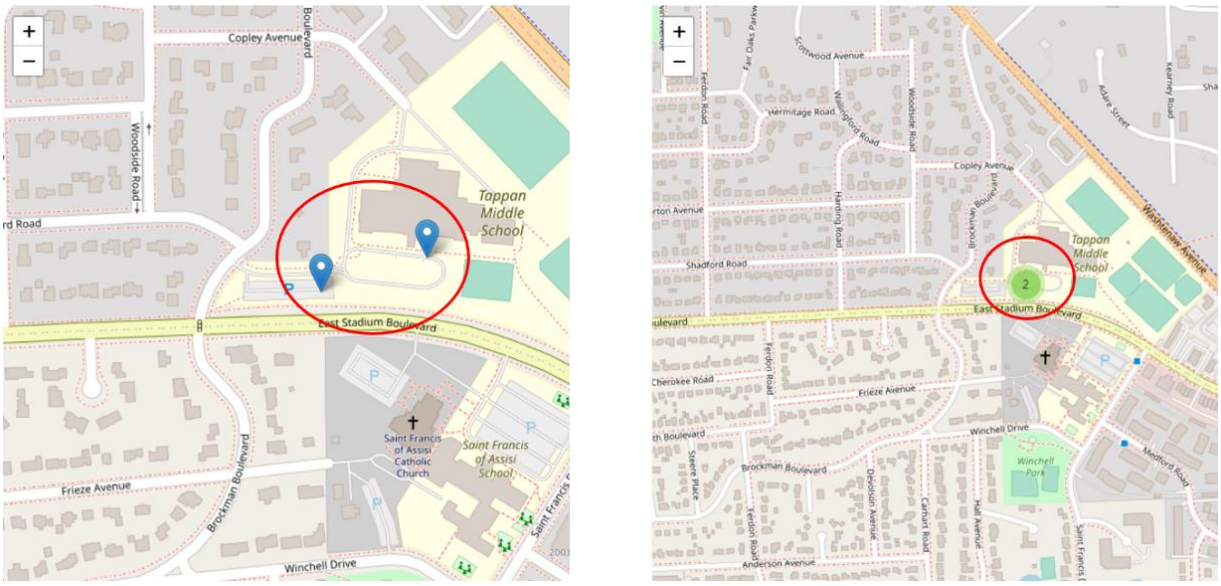


Figure 2.1 Sample trips from the telematic data collection survey

In this research, we represent a visited place through the coordinates (latitude and longitude) of the locations observed in the vehicle trajectory data. However, two different coordinates may represent the same visited place. For example, **Figure 2.2 (a)** shows two different visited coordinates of an individual, which are located within the same school campus. If we define these two coordinates as two separate locations, then we would not be able to capture the regularities in the activity pattern, because these two coordinates are essentially the same location with respect to an individual's activity participation. Thus, the two coordinates should be clustered into a group as shown in **Figure 2.2 (b)**. Such a cluster is also commonly referred to as a point of interest [35]. In this research, the visited coordinates are clustered based

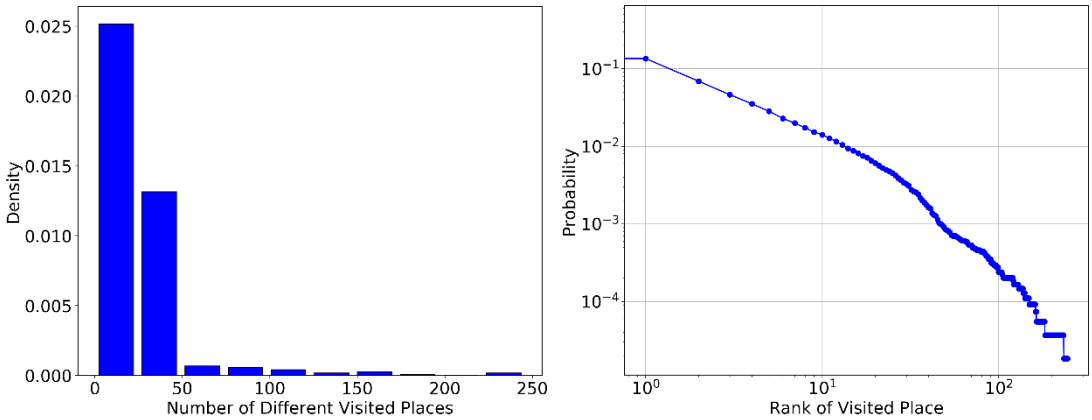
on a threshold so that if the distance between two visited coordinates is less than 200 meters, the coordinates are considered to be the part of a single point of interest.



(a) (b)

Figure 2.2 The visited coordinates (a) vs. visited place or point of interest (b)

The number of visited places varied across individuals as shown in **Figure 2.3 (a)**. It reveals that most individuals have visited fewer than 50 different places. Although an individual makes many visits, those visits concentrate on a limited number of locations, such as the home and work locations. This corroborates the findings from previous studies [45]. For each individual, we ranked her visited places based on the frequency of visits i.e., the most frequently visited place was ranked one, the 2nd most frequently visited place was ranked 2 and so on. **Figure 2.3 (b)** shows the probability of different ranks of the visited places across all the users. The first rank of the visited place accounts for more than 10% probability and the probability decreases sharply when the rank of the visited place increases. It also illustrates that when the rank of a visited place is more than 10, the probability that an individual would visit that place will be less than 1%, which means that an individual would seldom visit a low-ranked location.



(a) (b)

Figure 2.3 (a) The distribution of visited places number; (b) The probability of rank of visited places

In this research, we aim to predict the type and location of the next activity that an individual is going to participate in given some contextual information and the sequence of previous activities. To do this, we identify and label each activity participated by individuals based on the activity sequence. Since activity sequence can provide information on the type of visited places, the start time, and the activity duration, we can label activities using a set of rules. For instance, if a person starts an activity at 12 pm, continued the activity for 30 min, and the visited place was around a restaurant, then that activity is more likely to be an eating activity. The decision rules used to identify activities are listed in **Table 2.1**. The type of a place is defined based on the point of interest (POI) information data collected from the SafeGraph platform (https://www.safegraph.com/). From the SafeGraph platform, we extracted the POI coordinates located in Michigan. We extracted 151,561 POI records containing 147 categories such as restaurants and other eating places, supermarkets, malls, personal care services, and religious organizations.

Since the activity sequence of an individual is continuous, we consider that on each day, an individual will start his/her first activity from home and end the last activity at home. Thus,

among all the places visited by an individual, the most frequently last visited place is defined as the home place for that individual. In addition, for individuals with a regular job (they are expected to go to work on weekdays), we consider that the workplace will have the longest activity duration between 7 am and 5 pm and will be active at least 40% of the total active days (i.e., the days with GPS coordinates recorded). If the identified workplace is recorded in less than 40% of the total active days, the individual is defined as a non-working person.

The rest of the activities are labeled based on the start hour, duration, and the POI of the place. We label an activity when it satisfies all three requirements: start hour, POI information, and activity duration. For example, if an activity starts anytime between 11 am and 1 pm, the POI information shows that the activity happens near a food-related place, and the activity duration is 5 to 120 min, then the activity is identified as an eating activity. Otherwise, the activity cannot be identified as an eating activity and we will continue to infer the activity type. In terms of the POI information, since multiple POIs can be very close to each other, we do not choose the nearest POI of the activity location. For each activity location, we extract all the POIs which are within 200 meters of the specific activity location. When identifying an activity type, if the required POI information is in these POIs, we will consider that the activity location meets the requirement of the POI information.

To avoid having multiple labels for a visited place, we identify the activity type following a hierarchy using the level of the activity type. The home activity category is given the highest priority and the personal activity has the lowest. For example, if an activity is labeled as an eating activity, we will stop the identification process. Otherwise, we will check whether the activity is a shopping activity. If the activity cannot be identified as any of the 6 selected categories, we label the activity as ‘Not Identified’.

Table 2.1 Activity type identification rules

Level	Activity Type	Start Hour	POI information	Activity Duration (min)
0	Home	Any Time	Home location	Any Time
1	Work	Any Time	Work location	Any Time
2	Eating	11 – 13 hr 17 – 19 hr	Food, Restaurant…	5-120
3	Shopping	10 – 12 hr 14 – 22 hr	Supermarket, mall…	5-120
4	Pickup/Dropoff	Any Time	Any Place	< 10
5	Personal	5 – 22 hr	Gym, park, Hospital, dentist…	5- 240

One of the advantages of predicting activity behavior at an individual level is that we can not only predict the activity type, but also the specific activity locations. However, if activity locations are represented by GPS coordinates, they will be harder to predict. For example, two different coordinates may refer to the same visited places (home or work locations). To better predict activity locations, we divide the study area (Ann Arbor city, Michigan) into 10 by 10 grids. And each activity location is represented by a grid ID. **Figure 4** shows that each grid is given a grid ID between 0 and 99 and all the visited locations outside the study area have a grid ID 100. Each grid covers an area of about 0.42 square miles.

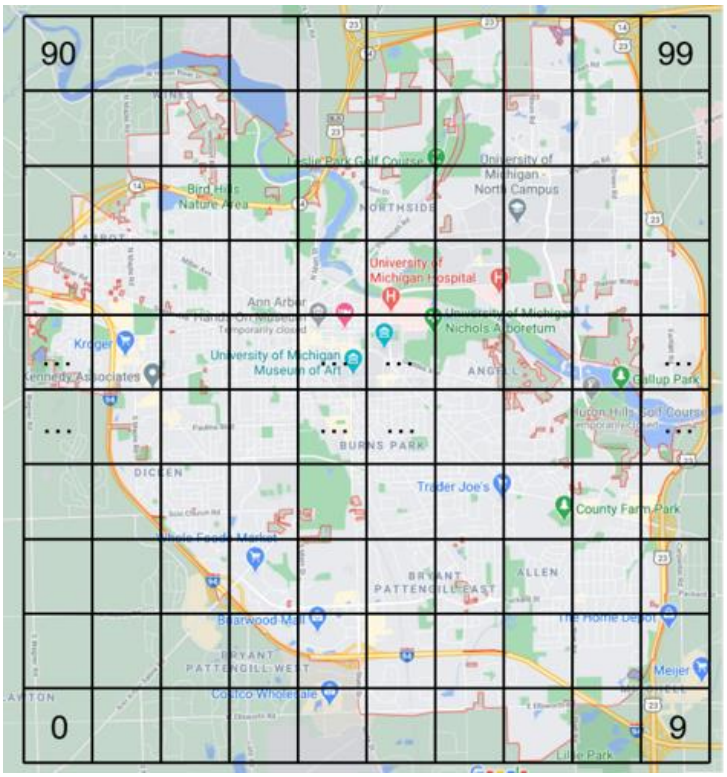


Figure 2.4 Grid division of the study area

2.4 Input-Output Hidden Markov Model (IOHMM)

Individual activity behaviors can be modeled from the historical trajectory sequence due to the presence of strong regularities as discovered from previous studies. However, the majority of the previous studies applied hidden Markov models (HMMs) to predict human movement. An HMM model uses homogeneous probability matrices for transition, and emission processes [69]. Since individual activity patterns have daily, weekly, and monthly periodicity, the HMM cannot identify all the periodicity with homogenous parameters. Thus, to better model individual activity sequences, we utilized an input-output hidden Markov model (IOHMM) which overcomes the shortcomings of the traditional HMMs.

Similar to the previous study [69], the framework of IOHMM adopted in this research contains three layers: input layer, hidden state layer, and output layer (observation layer) as shown in **Figure 5**. Compared to the traditional HMM, the IOHMM contains an additional input

layer that allows to introduce non-homogeneous parameters associated with different contexts of activity participation behavior. From **Figure 5**, it can be noted that the input layer affects both the hidden state, and the output layers and the hidden state layer influences only the output layer. To predict the next hidden state, the model will consider the information of the previous state and the corresponding input.

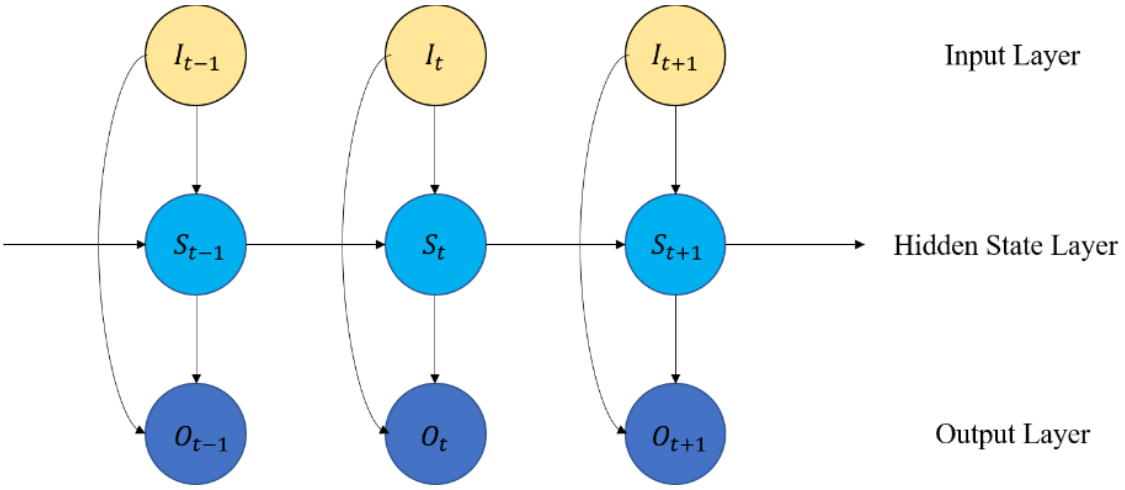


Figure 2.5 The framework of input-output hidden Markov model (IOHMM)

In the model, the input layer I_t provides the contextual information, such as time of the day (0 to 23 hr) and day of the week (Monday – Sunday). The IOHMM assumes that the contextual information contributes to the prediction of the hidden state which is behaviorally more plausible.

Since we cannot identify the types of all activities of a sequence, we use a semi - supervised learning algorithm with the hidden states as the targets while training the IOHMM. The hidden state layer S_t represents the activity type. Here, the number of hidden states is 6 representing six activity types: home, work, eating, shopping, pickup/drop off, and personal. The input layer I_t contains the contextual information, such as the time of the day (morning, lunch, afternoon, dinner, and evening), the day of the week (Monday, Tuesday, etc.), and whether the trip is the last trip of the day. The output layer O_t contains information on activity location.

In the standard HMM, the hidden states follow a Markov process with homogeneous transition and emission probabilities. It means that the current hidden state s_t depends only on the previous state s_{t-1}. Unlike the traditional HMM, in IOHMM, the current hidden state s_t depends both on previous state s_{t-1} and the input i_t at time t. For the framework of IOHMM, there are three main factors – initial probability, transition probability and emission probability. The initial probability decides the probability of the first hidden state s_1 given the first input i_1. The transition probability defines the relationship between the previous hidden state s_{t-1} and the current hidden state s_t given the current input i_t. The emission probability provides the probability of the output o_t given the current hidden state s_t and current input i_t.

We define the initial probability as:

$$p(s_1|i_1;\alpha_{in}) \tag{1}$$

The transition probability is defined as:

$$p(s_t = j|s_{t-1} = k; i_t, \alpha_{tr}) \tag{2}$$

The emission probability is defined as:

$$p(o_t|s_t = j; i_t, \alpha_{em}) \tag{3}$$

The likelihood of an activity sequence of the IOHMM can be calculated by:

$$L(\alpha, o, i) = \sum_s \left(p(s_1|i_1;\ \alpha_{in}) * \prod_{t=2}^{T} p(s_t|s_{t-1}, i_t;\ \alpha_{tr}) * \prod_{t=1}^{T} p(o_t|s_t, i_t;\ \alpha_{em}) \right) \tag{4}$$

In the equation, i, s, o are input variables, hidden states, and output variables, respectively; the $\alpha_{in}, \alpha_{tr}, \alpha_{em}$ are the parameters for initial, transition, and emission process, respectively.

To estimate the parameters, we utilized the commonly used Expectation-Maximization (EM) algorithm. The E-step calculates the log likelihood given the parameters and corresponding dataset and the M-step maximizes the log-likelihood by tuning the parameters. We run the EM algorithm using the code developed by [69].

For the initial layer, the variables are the time of the day and day of the week, which are discrete. Thus, we applied a multinomial logistic regression model to estimate the initial probability. If the number of hidden states is k, the initial probability can be calculated by the following Equation 5:

$$p(s_1 = j|i_1; \alpha_{in}) = \frac{e^{\alpha_{in}^j * i_1}}{\sum_k e^{\alpha_{in}^k * i_1}} \quad (5)$$

Where α_{in} represents the initial parameters and α_{in}^j is the initial parameters for the initial state being at state j.

For the transition layer, the hidden states are the activity types. Since the activity types are discrete variables. We used a multinomial logistic regression model to estimate the transition probability from the previous hidden state s_{t-1} to the current hidden state s_t. The transition process can be modeled as Equation 6:

$$p(s_t = j|s_{t-1} = q; i_t, \alpha_{tr}) = \frac{e^{\alpha_{tr}^{q,j} * i_t}}{\sum_k e^{\alpha_{tr}^{q,k} * i_t}} \quad (6)$$

Where α_{tr} represents the transition parameters and $\alpha_{tr}^{q,j}$ is the transition parameters for transitioning the hidden state q to hidden state j.

For the output layer, we have two output variables – activity duration and activity locations. The activity duration is a continuous variable, and we used a linear regression model for the output model. The output model can be written as Equation 7.

$$p\left({o_t}^d \middle| s_t = j; i_t, {\alpha^d}_{em}\right) = \frac{1}{\sqrt{2\pi}\sigma_j} \mathrm{e}^{-\frac{({o_t}^d - \alpha_{em}^{d,j} * i_t)^2}{2\sigma_j^2}} \quad (7)$$

Where $\alpha_{em}^{d,j}$ represents the emission parameters for the activity duration when the hidden state is j and σ_j is the standard deviation of the linear regression model.

For the output variable activity location, we used a multinomial logistic regression model for the outcome (Equation 8):

$$p\left({o_t}^l \middle| s_t = j; i_t, {\alpha^l}_{em}\right) = \frac{e^{\alpha_{em}^{l,j} * i_t}}{\sum_k e^{\alpha_{em}^{l,k} * i_t}} \quad (8)$$

Where $\alpha_{em}^{l,j}$ represents the emission parameters for the activity location when the hidden state is j.

To estimate the parameters of the IOHMM, we used the code shared by Yin et al. [69], which is available from https://github.com/Mogeng/IOHMM.

2.5 Results

2.5.1 Inferring activity type

To evaluate the performance of the activity type identification, we analyzed the results from three aspects: distribution of the start time of each activity type (**Figure 2.6**), distribution of activity duration (**Figure 2.7**), and activity sequence of an individual (**Figure 2.8**).

Figure 2.6 shows that the start time distribution of home activity peaks at 16 – 17 hr which is the time people finish their work. For the work activity, we can see most of the start time is during the morning peak hour which aligns with our expectations. Eating activity mainly

happens during lunchtime (11 – 13 hr) and dinnertime (17 – 19 hr). The shopping activities are conducted mostly after lunchtime and dinnertime. For pickup/drop off and personal activity, we can see that the distribution of the start time is quite random i.e., they can happen at any time of the day.

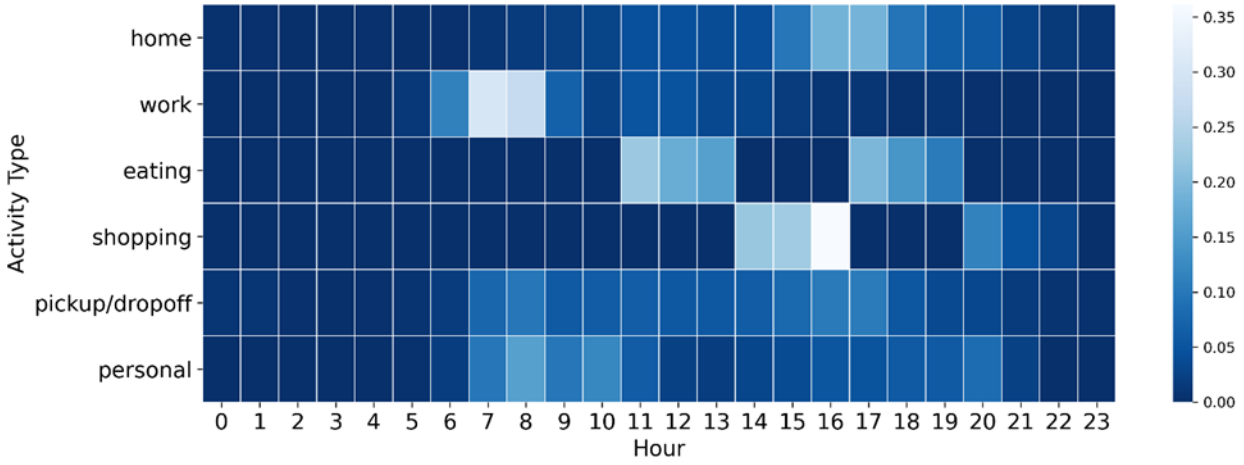


Figure 2.6 Distribution of start time for different activities

Figure 2.7 shows the distributions of activity duration by activity type. From the figure, we can find that the home and work activities have the largest duration – 8 hours peak for work and 7 hours peak for home (excluding overnight home activity). For eating and shopping activities, we can see most of them are less than 2.5 hours long. The pickup/drop off and personal activity have the shortest activity duration – less than one hour for personal activity and less than 10 min for pickup/drop off activity.

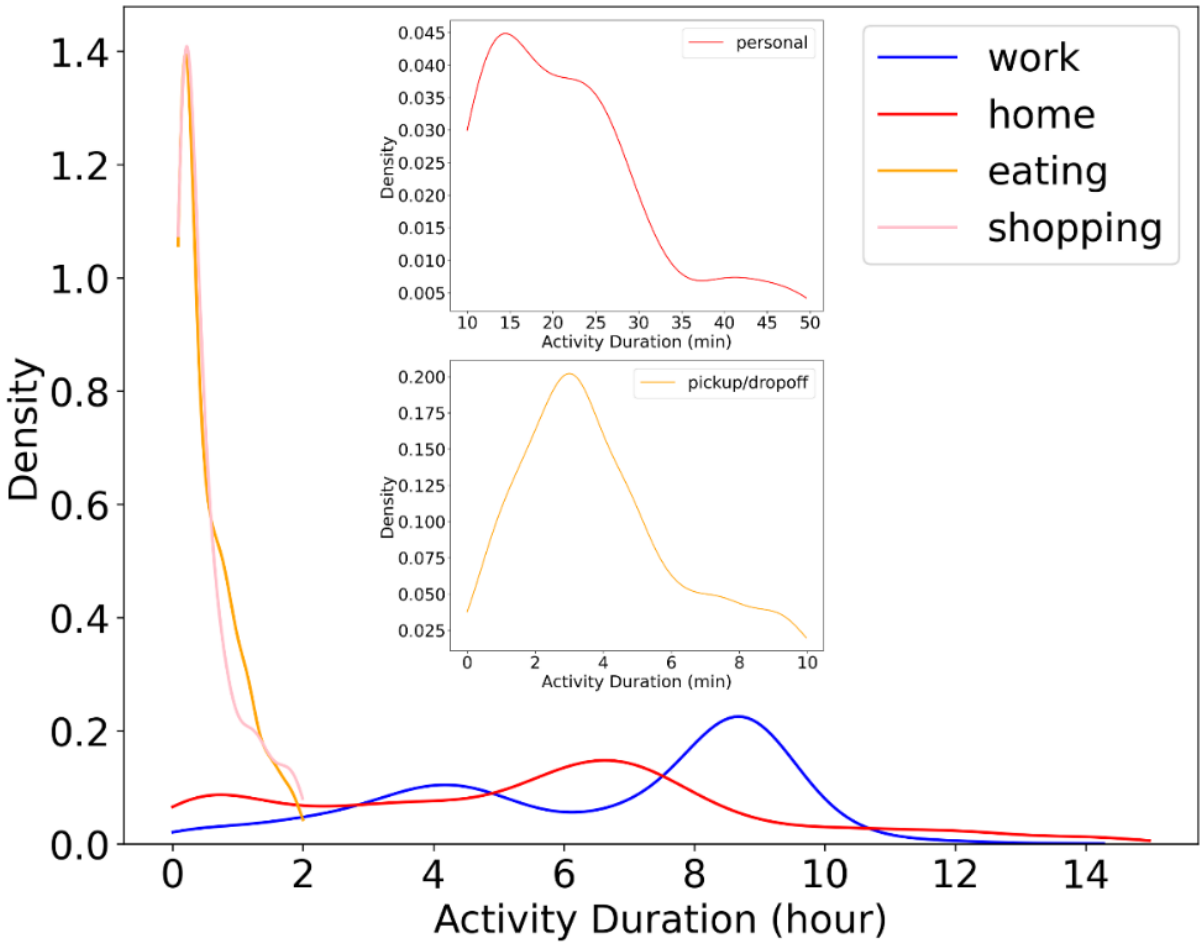
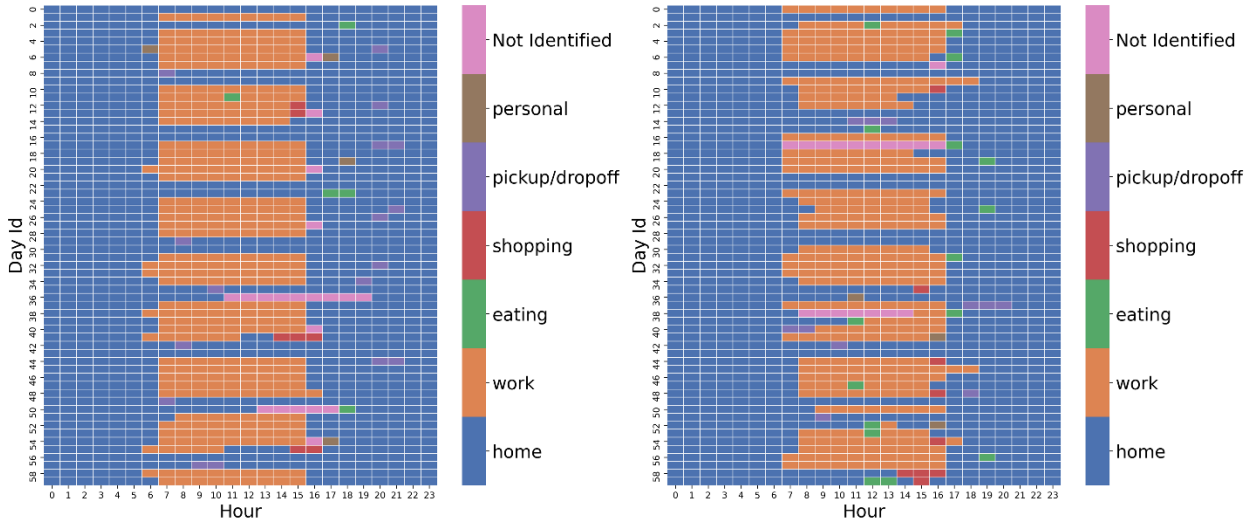


Figure 2.7 Activity duration distribution for each activity type

In addition to the start hour and the activity duration, we can generate individual-level daily activity sequence which is shown in **Figure 2.8**. We selected two randomly selected vehicles to generate the daily activity sequence plot with data covering 60 days. To present the activity sequence organized over multiple days, for each day, we assume that the last activity will be a home-based activity and the first activity in the next day will be a non-home-based activity; one needs to travel from home to participate to that activity. In the figure, the *y* axis indicates the day count, and the *x* axis presents the time of the day (in hour). Different colors in the figure represent distinct activity types, such as home, work, eating, and so on. As we can see from the two vehicles, we can easily identify the commuting patterns of each individual. On each

day, the first activity started at around 7 am, which was coded as a work activity by the algorithm. The green bars represent eating activities; some eating activities were identified during work hours while others were attended during the evening periods. These two users mostly participated in shopping activities right after work.



(a) (b)

Figure 2.8 Individual daily activity sequence with activity types for two vehicles (a) vehicle 9 and (b) vehicle 10

2.5.2 IOHMM model results

In this section, we analyze the results of the IOHMM for generating individual activity sequences. To train and evaluate the IOHMM, we applied the model to the trajectory data of 274 individuals. For each individual, we used 70% and 30% of activity sequences to train and evaluate model performance, respectively.

Since this is a generative model its prediction performance is evaluated first at an aggregate level. To do this, the generated number of activities of each activity type for different time periods are compared against the actual number of activities in the corresponding periods. We use mean absolute error (MAE) and mean absolute percentage error (MAPE) to evaluate the

performance of the developed generative model. The MAE and MAPE can be calculated by the following equations:

$$MAE = \sum_{i=1}^{t} \left| n_{pred}^{i} - n_{real}^{i} \right| \tag{8}$$

$$MAPE = \frac{\sum_{i=1}^{t} \left| n_{pred}^{i} - n_{real}^{i} \right|}{n_{total}^{i}} \tag{9}$$

Where i is different time periods (morning, lunch, afternoon, dinner and evening), n_{pred} is the number of predicted activities in the time period i, n_{real} is the actual number of activities in the time period i, and n_{total} is the total actual value of activity number in the time periods i.

We present the results of MAE and MAPE values in **Table 2.2**. It shows that for different time periods, the range of MAPE is from 9% to 18.7% which means that, in terms of activity number, our model can achieve accuracy levels between 81.3% and 91.0%. It also shows that the morning and lunch time periods have higher MAPE values, indicating that it is more difficult to correctly predict activities in the morning and lunch periods. The model can work better in the afternoon, dinner and evening periods, since the MAPE values in these three periods are around 10%.

Table 2.2 Results of comparison of activity number

Time Periods	Total Number of Activities	MAE	MAPE
Morning (5am – 10am)	22582	4214	18.66%
Lunch (11am – 1pm)	20243	3790	18.72%
Afternoon (2pm – 4pm)	22690	2494	10.99%
Dinner (5pm – 7pm)	18589	2236	12.03%
Evening (8pm – 12am)	7425	674	9.01%

To analyze the performance of the model for different activity types, we compare the generated and actual number of activities for each activity type as shown in **Figure 2.9**. It shows

that the model will generate more work activities in the morning period and more eating activities in the lunch period. Since holidays were included in selected data collection periods, the model would generate work activities while individuals may have participated in other activities (such as home or shopping activities) in the holiday periods. Besides, eating activities, which are done outside the home, will have more randomness. Both of these factors will generate more errors in the morning and lunch periods. For the other periods, we can see all the activities will have a similar number between the predicted value and the actual value.

We also compare the duration for each activity type in different time periods in **Figure 2.10.** The results show that the distributions of predicted activity duration are similar to that of the actual activity durations. For all the time periods, the distributions of activity duration have two peaks. In the morning periods, the first peak is the same for both the predicted value and the actual value which is about 50 minutes, and for the second peak, the predicted value is a little less than the actual value which may be due to more work activities generated. In the lunch periods, the two peaks of activity duration distribution are nearly the same for both the predicted value and the actual value. The first peak is between 50 minutes and 1 hour, and the second peak is about 650 minutes (probably home activity). For the afternoon periods, the predicted value and actual value also have similar distributions with two peaks (50 minutes and 500 minutes, separately). In the dinner activity, although the predicted value and actual value have similar peaks, the density of peaks of predicted value is higher than the actual value, since more eating activities are in the predicted value. In the evening period, there is only one peak which is about 200 minutes (most of the activities in the evening period are home activities).

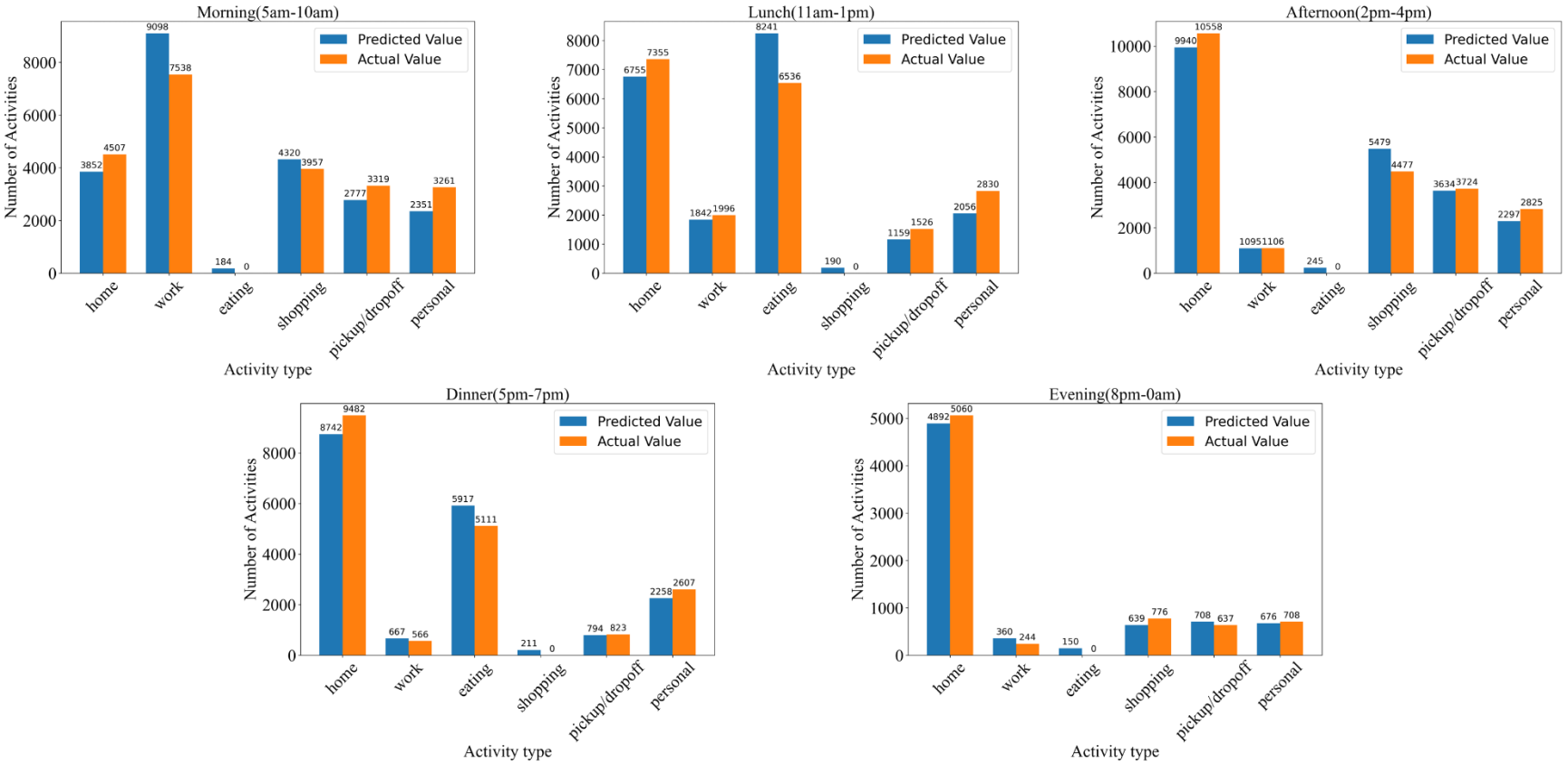


Figure 2.9 Results of number of generated and actual activities by type

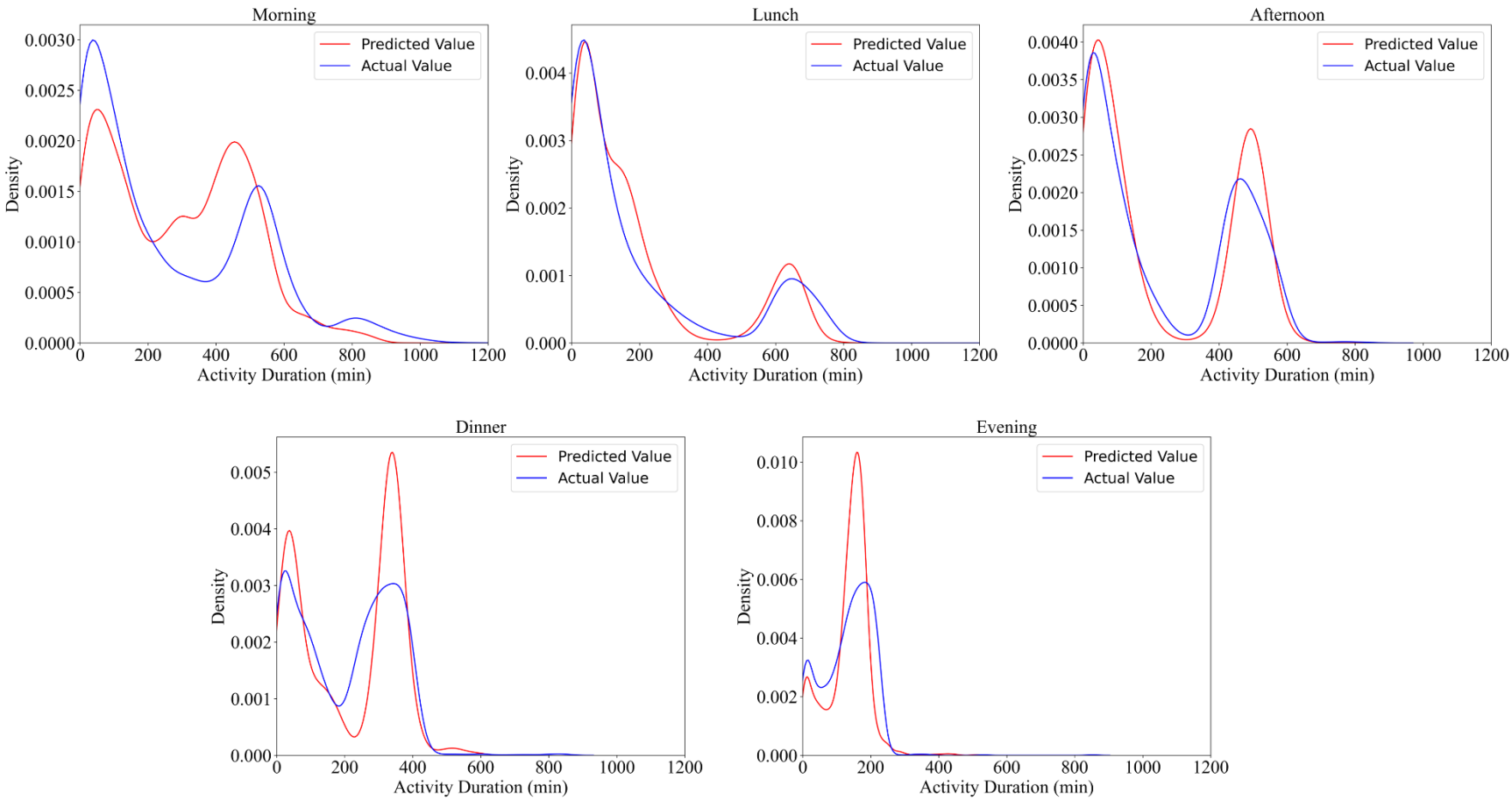


Figure 2.10 Results of number of generative activity duration vs. real activity types

When we know the type of a participated activity, we can further generate the activity location with contextual information and the results are presented in **Figure 2.11**. The developed model generated activity location in the output layer given the information of the input layer and hidden state layer. We evaluated the performance of activity location generation by calculating accuracy which is defined as the ratio of the number of correctly generated activity locations to the number of total activity locations. The activity location generation results show that the distribution of accuracy peaks at nearly 100% and more than 81% of individuals have at least 50% accuracy, which indicates the prediction model works well in predicting activity location. Some activities are conducted at fixed locations such as home activity at home and work activity as workplace which helped to attain higher prediction accuracy for activity location.

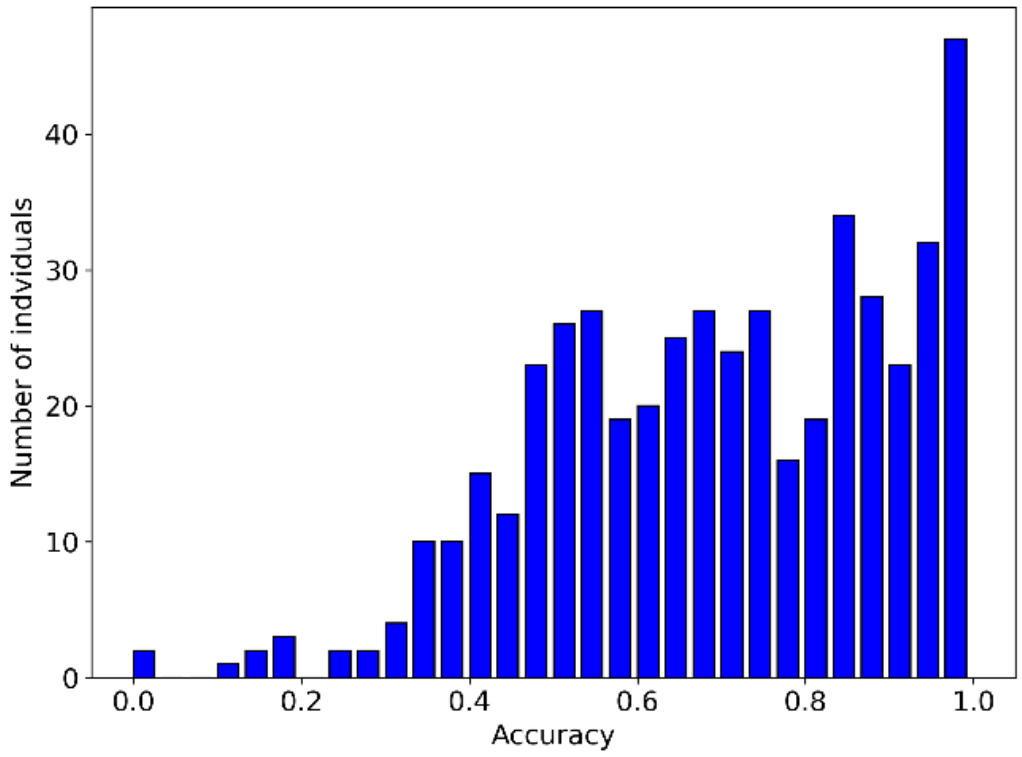


Figure 2.11 Distribution of model accuracy for activity location prediction

We show the initial probability distribution of each activity type for all the time periods in **Figure 2.12**. In the model, we assume that all the individuals will stay at home before making

a trip to participate in their first activity. Thus, the first activity of each individual for each day should not be a home activity. Due to this assumption the results show that the initial probabilities of home activity are 0 across all the time periods.

The results show that, in the morning period, work and shopping activities have the highest probability, followed by pickup/drop-off and personal activities. The initial probabilities of the other activities are near 0. It indicates that in the morning, individuals tend to go to work or shopping as the first activities and some individuals have pickup/drop-off and personal activities. In the lunch period, eating activity is the most popular among all the activities. Most people tend to have an eating activity as their first vehicle trip at lunch time. The shopping activity dominates the afternoon period. There are also some individuals who take work, pickup/drop-off and personal activities for them for trip of each day. In the afternoon, shopping activity is the most likely initial activity, which means that people prefer to participate in a shopping activity if they need to make a trip in the afternoon. If an individual takes the first vehicle trip, in the results, we can find that the work, eating, shopping, pickup/drop-off, and personal activities show a similar probability.

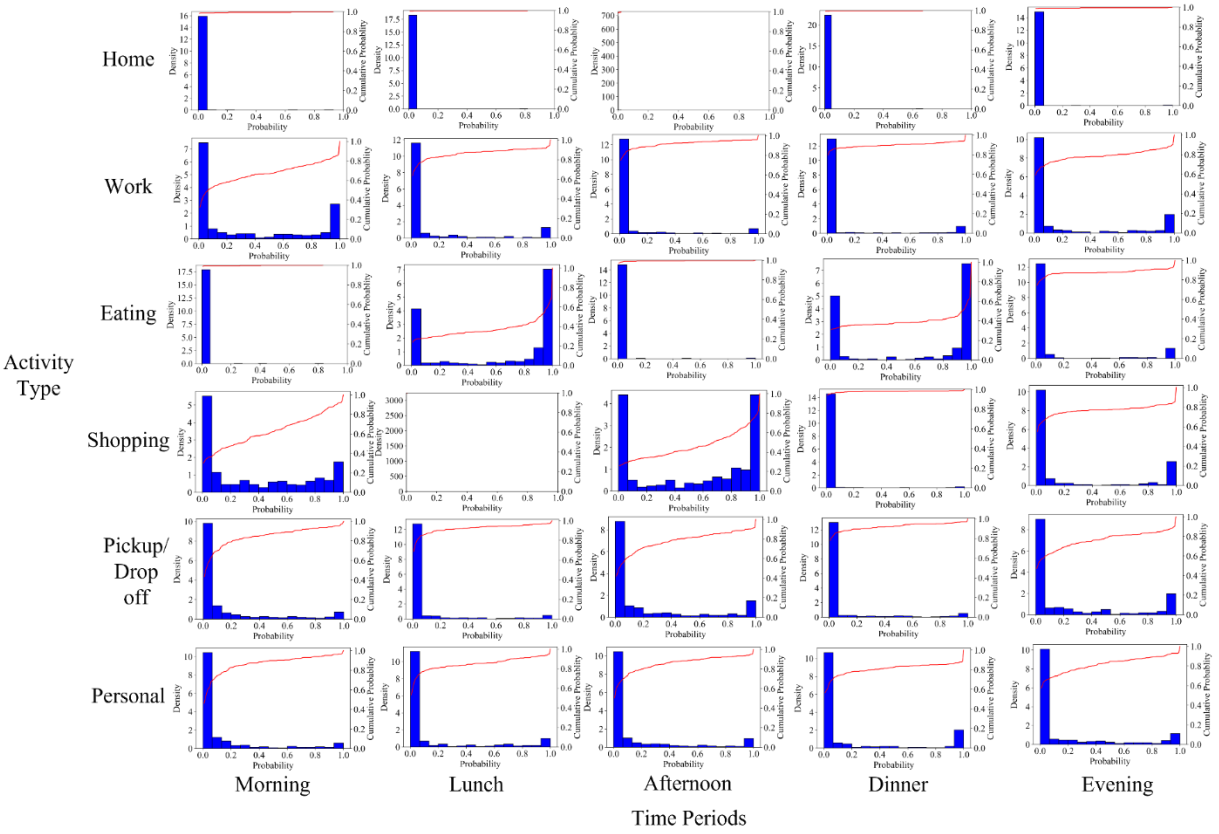
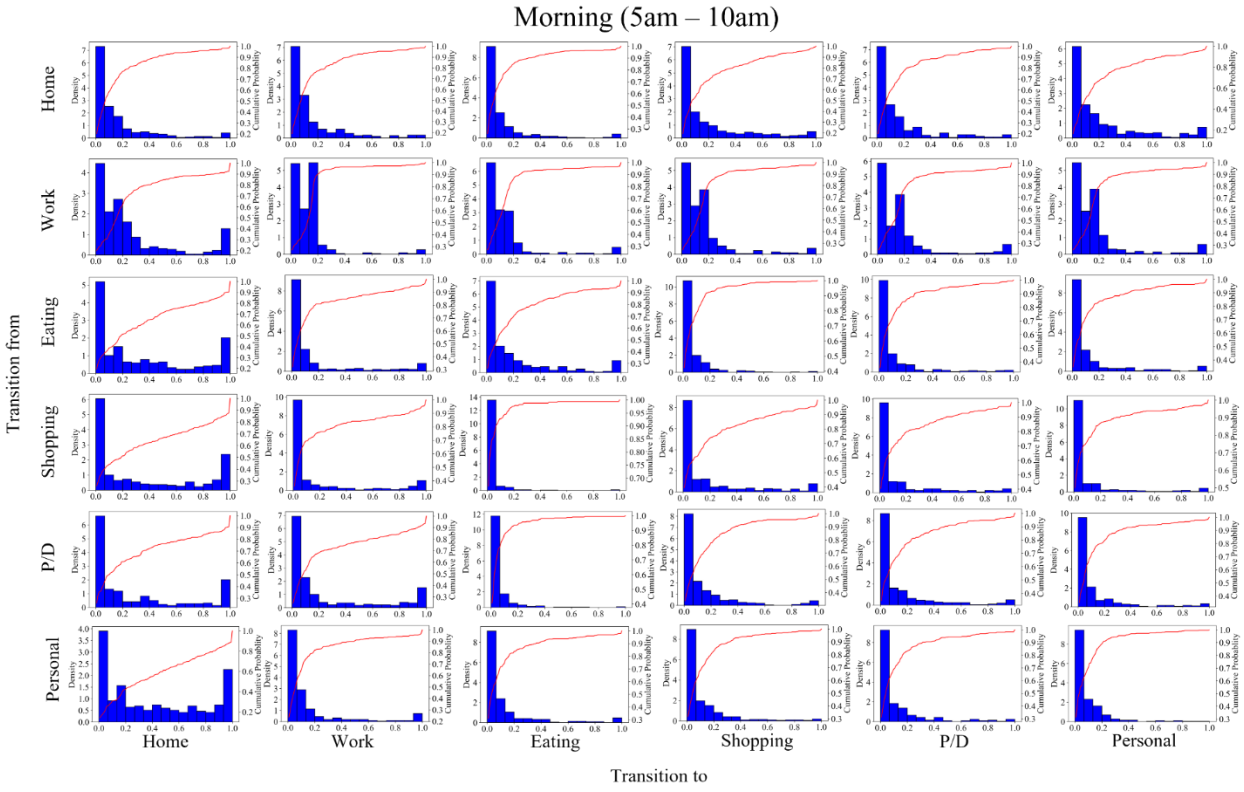


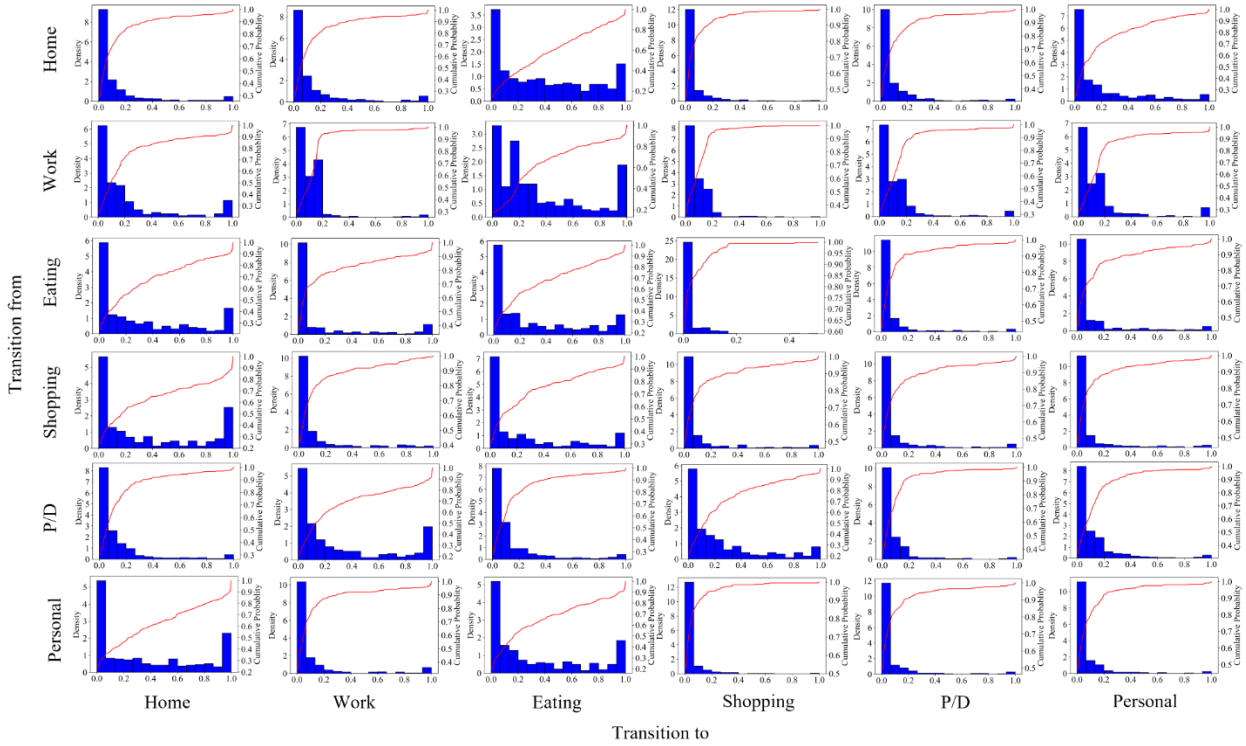
Figure 2.12 Distribution of initial probability in different time periods

We present the distribution of transition probability across all the periods for test individuals in **Figure 2.13**. For each period, we have a transition matrix with the distribution of transition probabilities. From the transition matrix associated with each time period, the labels on the left indicate the state the user is transitioning from and the labels on the bottom indicate the state the user is transitioning to. The most significant transitions are from the other activities to the home activity. Since it has been assumed that the first activity of each day cannot be a home activity, the first activity is more likely to be shopping and work activity, which will lower the probability of other activities transitioning to shopping and work activities. In the lunch time periods, individuals tend to transition from the home and work activities to eating activities. Besides, pickup/drop off activities are more likely to transition to work activity. In the afternoon time periods, there is a high probability for the users staying at home to transition to shopping

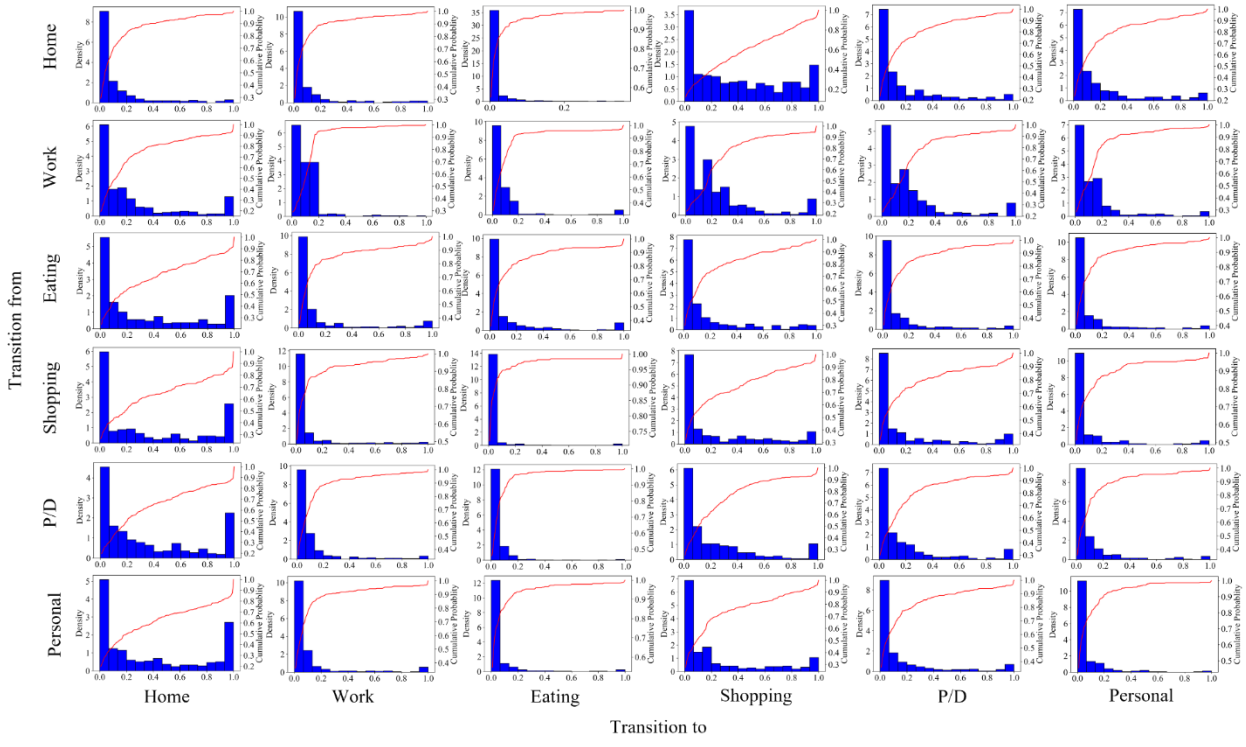
activity. And all the other states to home activity is significant. In the dinner period, patterns are similar to lunch period except for the pickup/drop off activity. Home and work activity will have a higher probability to transition to eating activity and users with all the other activities are more likely to go home. In the evening period, users tend to go home from all their other activities.



Lunch (11am – 1pm)



Afternoon (2pm – 4pm)



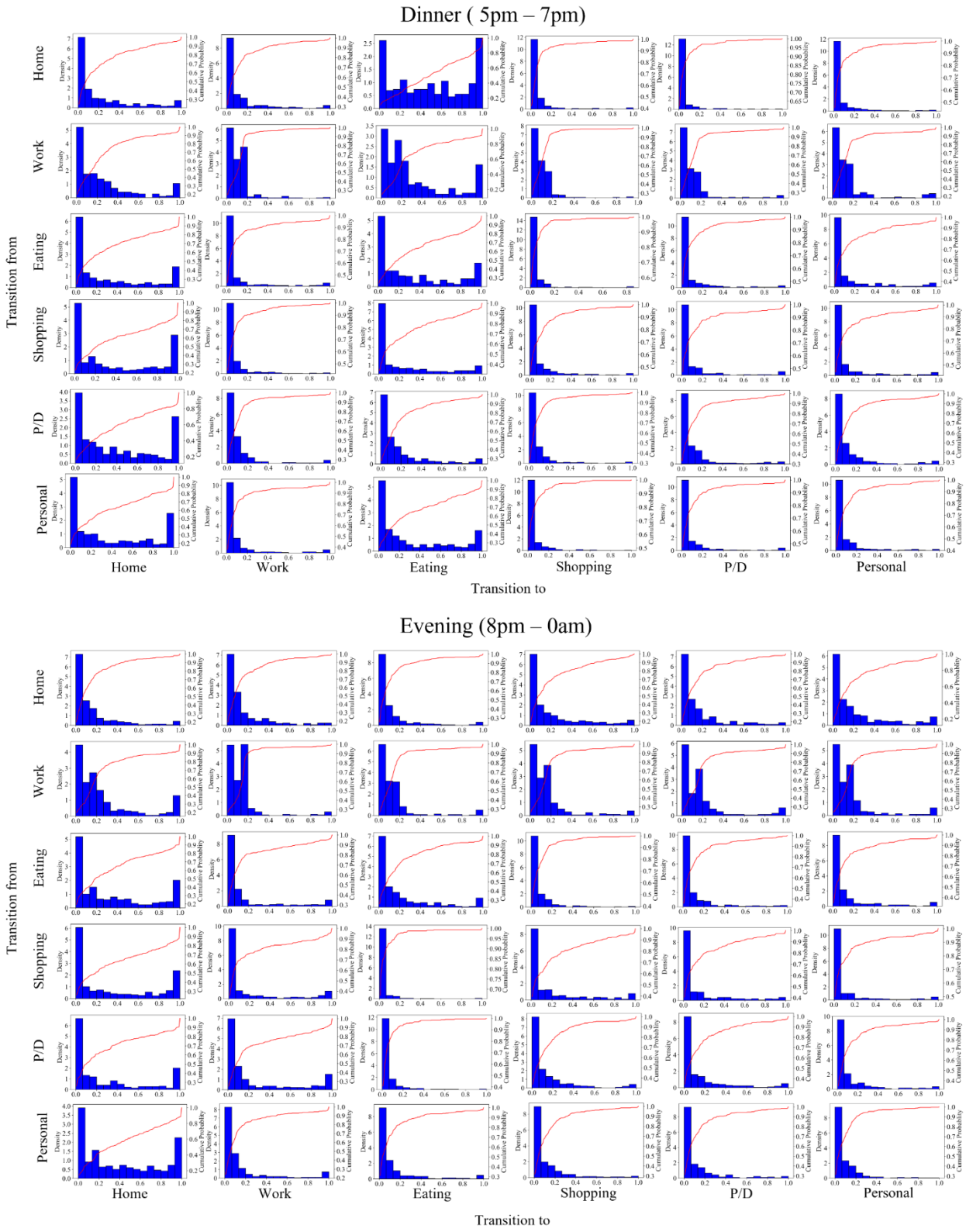
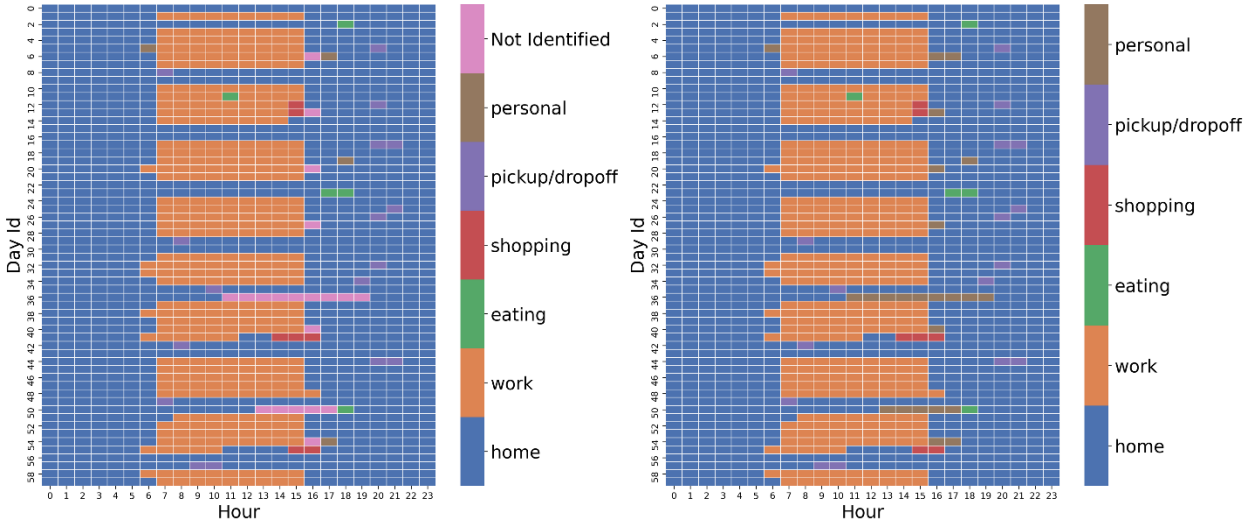


Figure 2.13 Distribution of transition probability in different time periods

One of the applications of our individual-level activity prediction model is that we can generate an activity sequence given the contextual information. We use vehicle 9 as an example to compare the ground truth of the activity sequence and the generated activity sequence. The results can be seen in **Figure 2.14**. First, we can find that in the ground truth of the vehicle 9 activity sequence, there are several activities that cannot be identified. In the generated activity sequence, the model can identify the corresponding activities based on the historical dataset which shows the potential of the individual activity prediction model in future applications. Besides, we can see the generated activity sequences are very similar to the ground truth activity sequence, which shows that although the general accuracy level is not high enough, the model can be used to generate realistic activity sequences for individuals.



(a) (b)

Figure 2.14 Ground truth of activity sequence vs. generated activity sequence: (a) ground truth of vehicle 9 activity sequence and (b) generated vehicle 9 activity sequence

2.6 Conclusions

In this paper, we present an input-output hidden Markov model (IOHMM) to generate individual activity sequences given contextual information. We first applied an algorithm to identify the

type of an activity based on POI category, start time of the activity, and duration of the activity. The results show that our algorithm works well on the activity type identification. We tested our model using massive individual vehicle trajectory data. To the best of our knowledge, the model used in this research is the first to predict individual activity patterns using full consecutive activity sequence available from vehicle trajectory data. Since the activity patterns for individuals are highly heterogeneous, we trained the IOHMM separately for each individual. From the comparison of activity number and activity duration between the predicted value and ground truth, it indicates the developed model can work well in predicting the activity sequence at an individual level. Besides, the results also show that if we know the activities of individual, the model can accurately predict the corresponding activity locations. To explain the model parameters, we analyze the distribution of initial and transition probabilities for different time periods.

One of the advantages of the proposed model is the explainability compared to the previous methods such as neural networks and other machine learning algorithms. From the model parameters, we can understand the transition probability for different activities considering contextual information such as the time of the day and day of the week. The explainable results can help us understand and validate models which can also be beneficial for improving the model accuracy through iteration. The proposed model can be used to generate the activity sequence for individuals given the contextual information. The results show that the generated activity sequence is very similar to the real activity sequence. In addition, the developed model can predict the activity of a sequence even if the types of activities are missing in the training data. The generated activity sequence shows the potential of the proposed model in generating realistic activity sequences at an individual level. These results show the great

potential of the developed model in real-world transportation planning applications such as traffic simulation, travel demand estimation, and OD matrix prediction.

There are also some limitations in this paper. First, although the proposed activity identification algorithm can achieve great performance by analyzing the activity's start time and duration, the ground truth data about activity participation will be required to validate the results of the algorithm. In the future, if more real-world individual activity data are available, the predictive model can be improved. Second, the developed model has a poor performance in predicting personal activities since we do not have socio demographic information such as age, gender, and the job of individuals.

CHAPTER 3: PREDICTING INDIVIUDAL MOBILITY BEHAVIOR OF RIDE-HAILING SERVICE USERS

3.1 Introduction

[1] Understanding individual mobility behavior can provide insights for transportation planning and traffic management. Based on the historical travel patterns, an individual's daily travel activities can be predicted. Combining individual travel patterns, we can forecast aggregate traffic demand so that transportation agencies can make strategies to mitigate congestion problems. To predict individual mobility behavior, researchers have made significant efforts using a wide variety of data sources. Traditionally, household travel surveys [71, 72] are the main data source to predict individual mobility behavior. However, high cost, low efficiency, and low sampling rates limit the application of travel surveys. Recently, data from banknotes [53], smartphones [42, 58], social media [28, 43, 44], and smart card transactions [25, 45] have been used to reveal spatio-temporal patterns of individual mobility. However, these data sources have limitations in modeling human movement. For instance, data from banknotes, smartphones and social media platforms have missing activities and they do not have information on the travel mode of each trip. On the other hand, data from public transport smart card transactions can provide the information of the travel mode (e.g., bus, train); but they do not offer the precise origin and destination of a trip since each trip starts/ends at a train station or a bus stop. Thus, to better understand individual mobility behavior, data with precise transportation related information (e.g., origin, destination, travel mode) are essential.

[1] Zhang, J., Hasan, S., Roy, K.C. and Yan, X., 2021, September. Predicting Individual Mobility Behavior of Ride-Hailing Service Users considering Heterogeneity of Trip Purposes. In 2021 IEEE International Intelligent Transportation Systems Conference (ITSC) (pp. 3685-3690). IEEE.

In recent years, on-demand ride-hailing services (such as Uber, Lyft, and Didi) have become emerging transportation modes in our daily life. These companies provide an online platform for individuals so that users can book a private ride-hailing service vehicle or a taxi for the next trip by giving the origin and destination of the trip. This innovative transportation platform can serve millions of users per day. The data from the ride-hailing services offer us a great opportunity to understand and model individual mobility behavior over many days. Although previous studies analyzed on-demand ride-hailing service data for travel demand prediction [73] and ride-splitting behavior analysis [74] at an aggregate level, there are few models predicting the travel behavior of on-demand ride-hailing service users at an individual level.

Besides, mobility behaviors of ride-hailing service users are prone to irregularity or randomness [21]. However, previous studies have seldom investigated the role of randomness for ride-hailing service users' mobility behavior. Thus, one important question remains to what degree the mobility behavior of ride-hailing service users is predictable. The predictability of mobility behavior can be also used to evaluate the performance of potential individual mobility prediction models.

In this study, we develop an innovative Markov chain-based model, multi-layer hidden Markov model, to predict individual mobility behaviors using the data extracted from Didi which is one of the largest on-demand ride-hailing platforms of the world. To validate the performance of the proposed model, we compare our results against a hidden Markov model (HMM). Besides, we also calculate the predictability based on Fano's inequality [75] to investigate whether the accuracy of individual mobility prediction model can achieve the theoretical limits of predictability of the mobility of ride-hailing service users.

The main contributions of this study can be summarized as follows:

- We develop an individual-level mobility model using a multi-layer hidden Markov based approach to predict both the trip making and origin and destination of ride-hailing service users. To the best of our knowledge, this is one of the few studies forecasting individual mobility behavior using massive ride-hailing service data.
- To consider the heterogeneity of mobility behavior among ride-hailing service users, we classify the users into four groups using their historical mobility characteristics and develop models to predict the mobility behavior of each group.
- We investigate the predictability of the mobility sequence of ride-hailing service users and validate the performance of trip origin prediction results using the concept of predictability of human movement.

3.2 Literature Review

Individual mobility behavior is one of the main elements of urban transportation systems modeling. Based on and emerging data sources (e.g., mobile phone, transit smart card, social media data), many studies analyzed individual movement patterns [23, 53, 54] or developed models for individual mobility behavior [35, 37, 71] using various modeling techniques. Recently researchers have developed statistical models to predict individual mobility behavior, such as Bayesian *n*-gram model [25] and Markov chain-based model [31, 58, 67, 69, 70, 76]. In addition, various machine learning based models such as neural networks [65] and genetic algorithms [35] have been developed.

Among all the models used for predicting individual mobility behavior, the Markov chain-based model attracts more attention. Using smartphone data, Xiu Lu et al. [58] developed a Markov chain model to predict the visited locations with a range of 87% - 95% accuracy, which

shows that individual mobility behaviors have strong regularities. Other studies [17, 27, 28][77] also reported that the Markov chain model can reach a range of 65% - 95 % accuracy with Wi-Fi wireless or GPS data of individual movement. Furthermore, in the study [37], the researchers applied an inhomogeneous continuous-time Markov model to predict the origin time and destination of the next trips of individuals with an accuracy of 67%. Besides, Qiujian Lv et al. [69] proposed a hidden Markov model to predict the individuals' points of interest using the data from cellular data networks. Differentiating with the previous studies, this study divided the individuals into distinct groups by their living habits based on entropy. The results show that the accuracy of the model varies from 20% to 70% with different time periods.

Human movement patterns reveal both regularities and uncertainties which will limit the predictability of mobility sequence. Researchers investigated the predictability of mobility sequence in recent years [21, 58, 62, 78]. In the study [21], using around 50,000 individuals' smartphone data, it reported a 93% predictability of individual mobility. However, the study did not investigate the relationship between predictability and the accuracy of prediction models. Using mobile phone data, Xin Lu et al. [58] showed that the predictability of more than 500,000 users' mobility behavior is around 88%. They also developed a Markov chain model showing a range of 87% - 95 % accuracy which indicates that the method of predictability can be the approachable goal for the real prediction accuracy.

Based on the aforementioned information, there is still a research gap in the literature to develop individual-level mobility models from ride-hailing service data and understand the predictability of user mobility behaviors. To fill this research gap, here we develop an innovative Markov chain-based model, an input-output hidden Markov model, to predict individual mobility

behavior and investigate the relationship between model accuracy and predictability of trips made by ride-hailing service users.

3.3 Data and Methods

3.3.1 Data Description

In this study, to develop the proposed prediction model, we used user-level mobility data extracted from the Didi ride-hailing service, which is the largest on-demand ride-hailing service platform in China serving more than 400 cities [74]. The dataset contains more than 5,000,000 on-demand ride-hailing users with the identification of trips and passengers, the coordinates of origins and destinations of each trip and trip attributes (travel time, travel distance and cost) from March 1, 2017 to June 31, 2017. The study region covers the area inside Beijing's 6th ring road (**Figure 3.1**). To reduce the computational cost, we randomly selected 50,000 users to develop the prediction model. Besides, we also collected the weather information from the National Weather Science Data Center [79], since previous studies [32, 77] reported that weather may affect individual mobility behavior. We divided the weather information into two fields – weather and air quality. The details can be seen in **Table 3.1**.

Table 3.1 Detailed Data Attributes

Mobility Data	
Field Name	**Field Description**
Record ID	The encrypted record id of one trip
Passenger ID	The encrypted passenger id of one trip
Driver ID	The encrypted driver id of on trip
Longitude of Origin	The longitude of the origin
Latitude of Origin	The latitude of the origin
Longitude of Destination	The longitude of the destination
Latitude of Destination	The latitude of the destination
Start Time	The timestamp of the origin
Arrive Time	The timestamp of the destination
Travel Distance	The travel distance of the trip (km)
Cost	The price of the trip record (RMB)
Weather Data	
Field Name	**Field Description**
Weather	0: no rain; 1: slight rain; 2: heavy rain
Air Quality	0: no air pollution; 1: slight air pollution; 2: heavy air pollution

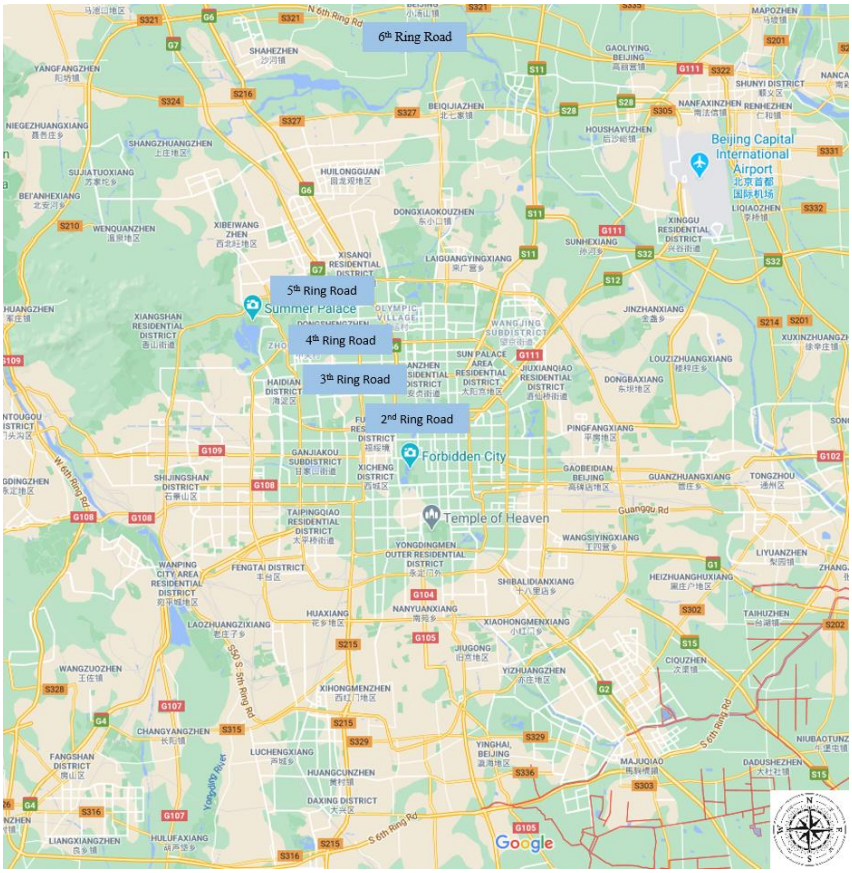


Figure 3.1 The study region: area inside Beijing 6th ring road

3.3.2 Data Exploration

Figure 3.2 shows some trip-related metrics of the selected 50,000 users. We find that that most of the selected users have made about 50 trips in four months (see **Figure 3.2 (a)**). Since not all the users have a trip every day, we also investigated the distribution of active days for individuals. **Figure 3.2 (b)** demonstrates that during the four months, the majority of the users were active in around 30 days. Besides, we also found that some users have less than 10 active days revealing that these users may be visitors. Since our study aims to find the typical mobility patterns, we removed the individuals who have less than 10 active days in the dataset. After removing the visitors, our final dataset contains 34,311 users.

We convert the coordinates of the trip origins and destinations as clusters, because sometimes two different coordinates do not indicate two different places. In terms of the historical origins and destinations of each individual, we defined that if the distance between two coordinates (either origins or destinations) is less than 300 meters, the two coordinates will be clustered as one visited place. **Figure 3.2 (c)** shows that most of the individuals have visited less than 20 different places by using the ride-hailing service in four months.

After we identify all the visited places for each individual, we also sort each one's visited places by their frequency and create a new variable called the rank of a visited place to represent each identified visited place. We also checked the distribution of the rank of visited places on a log-log scale in **Figure 3.2 (d)** for individuals with different numbers of visited places. It reveals that the top two ranks of visited locations have similar probability, but from the third ranked visited place and onward, the distribution decays following a Zipf's law. Similar pattern was observed for public transportation users in a previous study [45]. For individuals with a smaller number of visited places, the probability of the top two ranked locations will be higher. We can also find the probability that individuals travel to a higher rank (more than 10) visited place is nearly 1%. Thus, when developing the model, we limited the rank of the places up to 11, where all ranks higher than 10 are considered as 11.

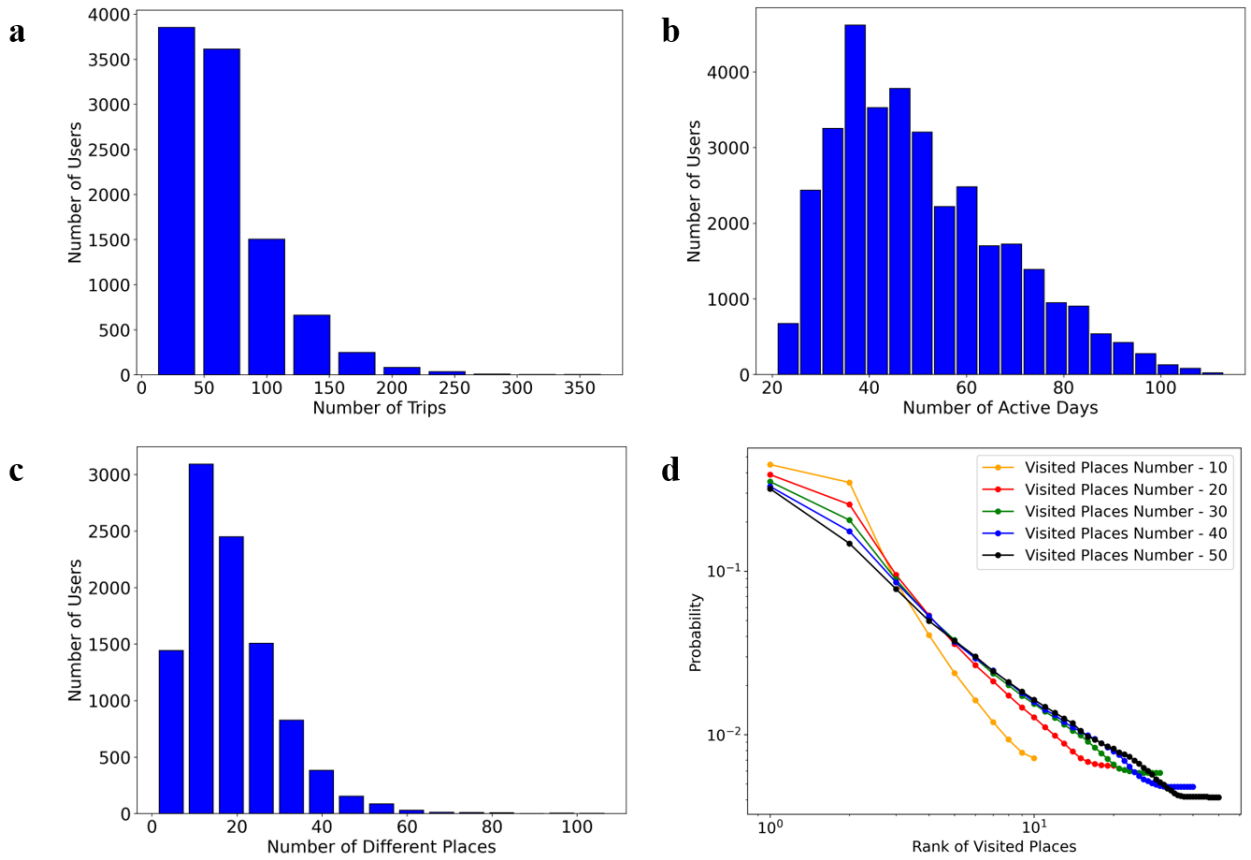


Figure 3.2 Exploratory data analysis: (a) the distribution of the number of trips; (b) the distribution of the number of active days; (c) the distribution of the number of visited places; (d) the distribution of rank of visited places in log-log scale

3.3.3 Clustering Users

From the dataset, we found the travel purpose for using the ride-hailing service is different across all the users. For example, some users utilize service vehicles only to come back home from his/her workplace and some users may use ride-hailing services for visiting different places for work purposes only. Thus, to better model the mobility patterns of ride-hailing service users, we cluster individuals based on their mobility habits. To cluster users, we mainly focus on their commuting patterns. To characterize individual commuting characteristics and to determine whether historical ride-hailing service data can reveal someone's home or workplace, we utilized two heuristic rules:

Home: For individual user, for a visited place j, we define the number of trips started from j in the extended morning peak hour (6 am – 11 am) as q_s^j; the number of trips ended at j in the extended evening peak hours (3 pm – 8 pm) as q_e^j, the total number of trips of individual is q_{total}.

For an individual, if $q_s^j + q_e^j$ is higher than the similar value for other visited places, the visited place j will have a higher probability to be the home place. To be sure, in this study, if $\frac{\max_j(q_s^j+q_e^j)}{q_{total}} > 0.4$, which means if the ratio of the largest sum of q_s^j and q_e^j to the total number of trips is more than 40%, then location j is defined as the home place. Otherwise, the individual's home place cannot be identified.

Workplace: For individual user, for a visited place k, we define the number of trips ended at k in the extended morning peak hour (6 am – 11 am) as q_e^k; the number of trips started from k in the extended evening peak hours (3 pm – 8 pm) as q_s^k, the total number of trips of individual is q_{total}.

Likewise, for individual i, if $q_s^k + q_e^k$ is higher than the similar value for other visited places, then the visited place k will have a higher probability to be the workplace. In this study, if $\frac{\max_k(q_s^k+q_e^k)}{q_{total}} > 0.25$, which means that if the ratio of the largest sum of q_s^k and q_e^k to the total number of trips is more than 25%, then, the location k is defined as the workplace. Otherwise, the individual's workplace cannot be identified.

According to whether the users' home and workplace can be identified, we clustered the ride-hailing users into four groups:

1) *Home-based users:* we can identify only the home of a user, which means that these users mainly use ride-hailing service for home-based trips.

2) *Work-based users:* we can identify only the workplace, which means that these users mainly use ride-hailing service for work-based trips.

3) *Commute-based users:* we can identify both home and workplace, which means that these users mainly use ride-hailing services for commuting purposes.

4) *Random users:* we can identify neither home nor workplace, which means that these users mainly use ride-hailing service for other purposes (neither home- nor work-based trips).

To show these four groups, **Figure 3.3** presents the hourly distribution of the number of origins by rank of visited places for a randomly selected user from each cluster. For instance, **Figure 3.3(a)** depicts a home-based individual mobility behavior, and we can see the user frequently starts from the rank 0 visited place in the morning peak hours, which represents that the rank 0 visited place is the home place of this individual. However, we cannot identify his/her workplace since he/she does not take many ride-hailing services in the evening peak hours. **Figure 3.3(b) – 3.3(d)** respectively show the samples of spatio-temporal mobility patterns of work-based users, commute-based users, and random users. From selected passengers, 6100 users are home-based, 4985 users are work-based, 11048 users are commute-based, and the remaining 12,178 users are random users.

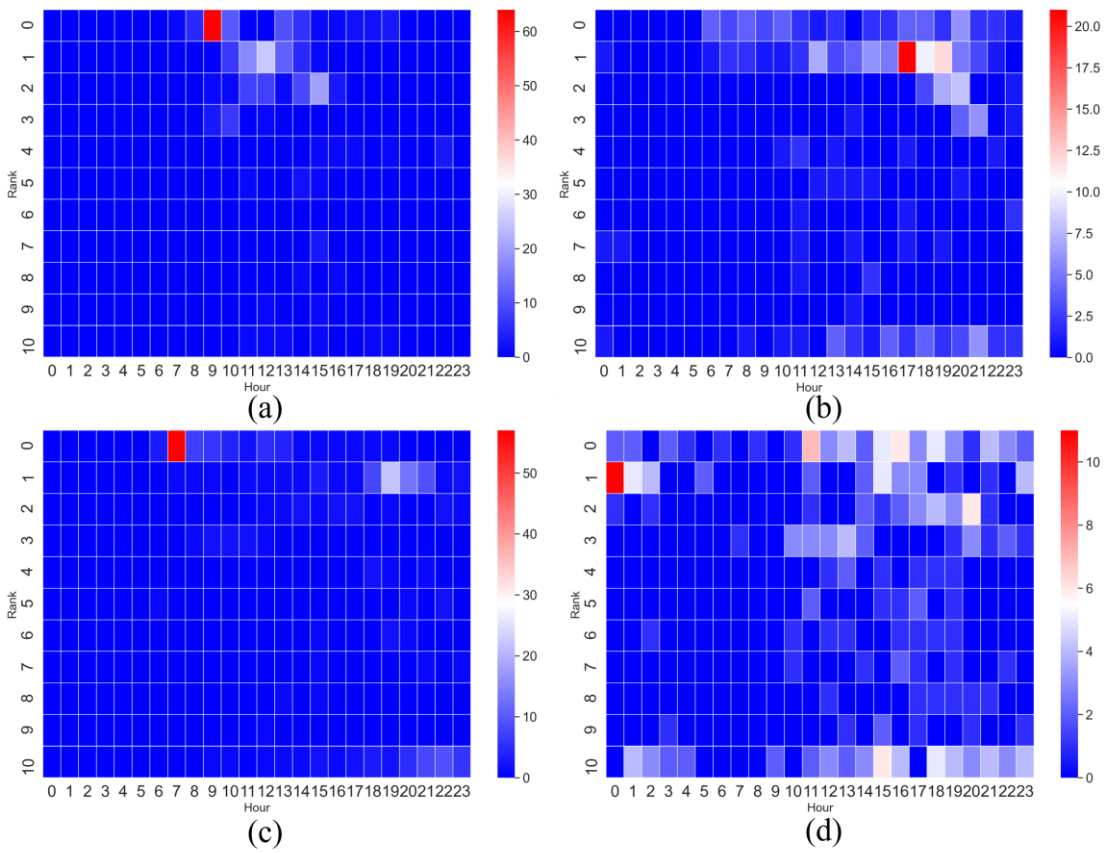


Figure 3.3 The distribution of rank of visited places in different hour: (a) home-based users; (b) work-based users; (c) commute-based users; (d) random users.

3.3.4 Multi-Layer Hidden Markov Model

A user may not use ride-hailing service every day. Thus, to better model the mobility behavior of on-demand ride-haling service users, one needs to predict the decision to take a ride-hailing service for a specific day and the characteristics of the ride-haling trips to be made in that day. The trip decision of the users represents whether the user will make a trip given the contextual information, and the characteristics of a ride-hailing trip will contain both the trip origin and destination. As such, the method used in this research is a multi-layer hidden Markov model which contains two parts: i) trip decision model and ii) mobility sequence generation model. We first use the trip decision model to predict whether the user will have a trip and then, if the user

has a trip, we will use the mobility sequence generation model to predict the ride-hailing service trips on that day.

Figure 3.4 shows the integrated modeling framework of trip decision and mobility sequence generation models. The trip decision model uses an input-output hidden Markov model (IOHMM) that includes three layers: input layer, hidden state layer, and output layer. The input layer (I) contains the contextual information such as the temporal information (day of the week), weather information, and individual characteristics (such as travel frequency level, radius of gyration, average travel cost and so on) extracted from the historical mobility sequence. The hidden state layer (S) will be the trip decision (whether to have a trip given the contextual information). The output layer (O) will be the number of trips on the specific day given the information of contextual and hidden state. Since the individual mobility sequence has weekly periodicity, for each individual, we process the data as a weekly sequence.

We define the initial probability of trip decision model as:

$$p(s_1|i_1;\alpha_{in}) \tag{9}$$

The transition probability of trip decision model is defined as:

$$p(s_t = j|s_{t-1} = k, i_t;\alpha_{tr}) \tag{10}$$

The emission probability of trip decision model is defined as:

$$p(o_t|s_t = j, i_t;\alpha_{em}) \tag{11}$$

The likelihood of the sequence of a trip decision model can be estimated by:

$$L(\alpha, o, i) = \sum_s(\Pr(s_1|i_1;\ \alpha_{in}) * \prod_{t=2}^{T}\Pr(s_t|s_{t-1}, i_t;\ \alpha_{tr}) * \prod_{t=1}^{T}\Pr(o_t|s_t, i_t;\ \alpha_{em})) \tag{12}$$

In the equation, i, s, o are input variables, hidden states, and output variables, respectively; the $\alpha_{in}, \alpha_{tr}, \alpha_{em}$ are the parameters for initial, transition, and emission processes, respectively.

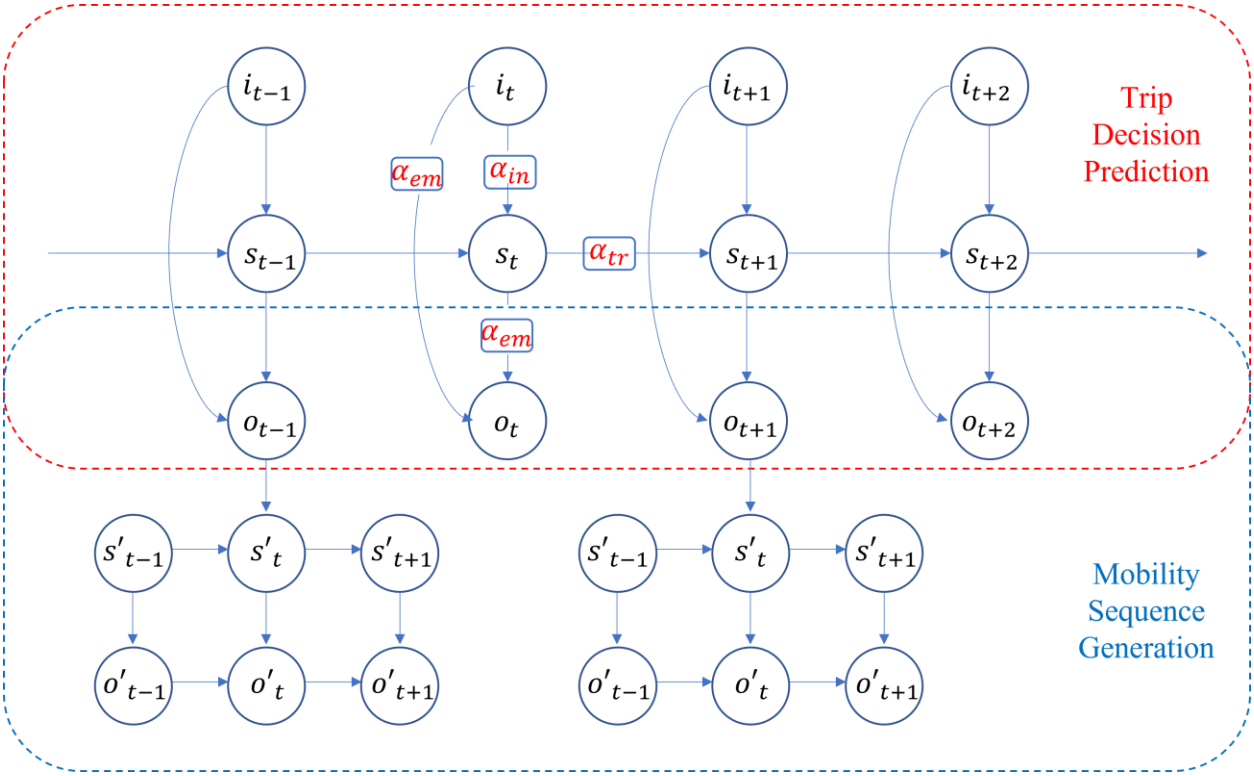


Figure 3.4 Multi-layer hidden Markov model framework

The second part of the model will be the mobility sequence generation model, which uses the results of the trip decision model. When the trip decision model shows that there will be a trip under a specific context, the mobility sequence generation model will generate the ride-hailing service trips (including the origin and destination of the trips) of the day. Instead of using the actual coordinates or stations of the visited places [25], we use the rank of a visited place as a target variable in the model. Similar approach has been used in previous research [80].

The mobility sequence generation model also uses an input-output hidden Markov model framework. The input contains two types of information – the contextual information (I) and the number of trips (O). The hidden state layer (I') will be the origin of each trip and the output layer (O') will be the destination of each trip.

We define the initial probability of mobility sequence generation model as:

$$p(s'_1|i_1, o_1; \alpha'_{in}) \tag{13}$$

The transition probability of mobility sequence generation model is defined as:

$$p(s'_t = j|s'_{t-1} = k, i_t, o_t; \alpha'_{tr}) \tag{14}$$

The emission probability of mobility sequence generation model is defined as:

$$p(o'_t|s'_t = j, i_t, o_t; \alpha'_{em}) \tag{15}$$

The likelihood of the sequence of mobility sequence generation model can be estimated by:

$$L(\alpha, o, i) = \sum_s (\Pr(s'_1|i_1, o_1;\ \alpha'_{in}) * \prod_{t=2}^{T} \Pr(s'_t|s'_{t-1}, i_t, o_t;\ \alpha'_{tr}) * \prod_{t=1}^{T} \Pr(o'_t|s'_t = j, i_t, o_t;\ \alpha'_{em})) \tag{16}$$

In the equation, $i,\ s', o'$ are input variables, hidden states, and output variables, respectively of a mobility sequence generation model; the $\alpha'_{in}, \alpha'_{tr}, \alpha'_{em}$ are the parameters for initial, transition, and emission processes, respectively, of the mobility sequence generation model.

The model selection depends on the type of variables. For example, in the trip decision model, the hidden state variables are discrete. Thus, to specify the initial and transition processes of the trip decision models, we use multinomial logistic regression models (Equation 17 and 18). We define the emission model with a linear regression model (Equation 19) to determine the number of daily trips.

$$p(s_1|i_1, o_1; \alpha_{in}) = \frac{e^{\alpha_{in} * i_1 * o_1}}{\sum_k e^{\alpha_k * i_1 * o_1}} \tag{17}$$

$$p(s_t = j|s_{t-1} = k, i_t, o_t; \alpha'_{tr}) = \frac{e^{\alpha_k^j * i_t * o_t}}{\sum_n e^{\alpha_k^n * i_t * o_t}} \tag{18}$$

$$p(o_t|s_t = j, i_t, o_t; \alpha_{em}) = \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\frac{(o_t - \alpha_{em} * i_t * o_t)^2}{2{\sigma_j}^2}} \tag{19}$$

The α_k^j represents the transition probability from the state k to the next state j and σ_j is the standard deviation of the linear regression model when the hidden state is j.

In addition, in the mobility sequence generation model, both the hidden state variable (the origin of a trip) and the emission variable (the destination of a trip) are discrete. Thus, all of the initial, transition and emission models are defined as multinomial logistic regression models shown in Equation 20 – 22, respectively.

$$p(s'_1|i_1;\alpha'_{in}) = \frac{e^{\alpha'_{in}*i_1*o_t}}{\sum_k e^{\alpha'_k*i_1*o_t}} \tag{20}$$

$$p(s'_t = j|s'_{t-1} = k, i_t, o_t; \alpha'_{tr}) = \frac{e^{\alpha'^j_k*i_t*o_t}}{\sum_n e^{\alpha'^n_k*i_t*o_t}} \tag{21}$$

$$p(o'_t|s'_t = j, i_t, o_t; \alpha'_{em}) = \frac{e^{\alpha'^j_{em}*i_t*o_t}}{\sum_n e^{\alpha'^n_{em}*i_{t*o_t}}} \tag{22}$$

To train the IOHMM, we used a supervised learning algorithm with the hidden states as the targets. Due to the heterogeneity in individual mobility behaviors, we train the mobility sequence generation model for home-based users, work-based users, commute-based users, and random users separately. For each group, we combined 70% of the historical mobility sequence of each individual to create the training dataset and the remaining 30% of the mobility sequence of each individual as the test dataset. To evaluate the performance of the proposed prediction models, we utilized accuracy as a metric. For each target variable (trip decision, origin rank, or destination rank), the accuracy is defined by the ratio of accurate predictions to all predictions.

To estimate the parameters, we utilized an Expectation-Maximization (EM) algorithm [81] which is commonly used for parameter estimation of hidden-Markov based models. The E-step calculates the log likelihood given the parameters and corresponding dataset and the M-step

maximizes the log-likelihood by tuning the parameters. We adopted the EM algorithm following the descriptions given in [26, 69]:

In the E-step, the estimated parameters at iteration $w-1$ are defined as α^{w-1}. If $w=1$, then the initial parameters are used. With α^{k-1}, we can obtain the initial, transition and emission probabilities, defined as π_j^{w-1}, $\varphi_{j,k;t}^{w-1}$, and $\delta_{j;t}^{w-1}$. Then, the forward ($\beta_{j;t}^{w}$) and backward ($\gamma_{j;t}^{w}$) variables can be calculated as:

$$\beta_{j;t}^{w} = \delta_{j;t}^{w-1} \sum_{l \epsilon S} \varphi_{l,j;t}^{w-1} * \beta_{l;t-1}^{w} \tag{23}$$

$$\gamma_{j;t}^{w} = \sum_{l \epsilon S} \varphi_{j,l;t}^{w-1} * \gamma_{l;t+1}^{w} * \delta_{l;t+1}^{w-1} \tag{24}$$

With the forward and backward variables, we can calculate the posterior state probability $\varepsilon_{j;t}^{w}$ and posterior transition probability $\theta_{j,k;t}^{w}$ as:

$$\varepsilon_{j;t}^{w} = \beta_{j;t}^{w} * \gamma_{j;t}^{w} / L_c^{w} \tag{25}$$

$$\theta_{j,k;t}^{w} = \varphi_{j,k;t}^{w-1} * \beta_{j;t-1}^{w} * \gamma_{k;t}^{w} * \delta_{k;t}^{w-1} / L_c^{w} \tag{26}$$

where L_c^{w} is the complete log likelihood at iteration w, which can be calculated by $\sum_{j \epsilon S} \beta_{j;T}^{w}$.

In the M-step, the parameters in iteration w are updated by maximizing the expected log likelihood:

$$Q(\alpha; \alpha^{w-1}) = \sum_{j \epsilon S} \varepsilon_{j;1}^{w} * logP(s_1 | i_1; \alpha_{in}) +$$

$$\sum_{t=2}^{T} \sum_{j,k \epsilon S} \theta_{j,k;t}^{w} * logP(s_t = j | s_{t-1} = k, i_t; \alpha_{tr}) + \sum_{t=1}^{T} \sum_{j \epsilon S} \varepsilon_{j;t}^{w} * logP(o_t | s_t = j, i_t; \alpha_{em}) \tag{27}$$

Then, we can obtain $\alpha^{w} = argmax_{\alpha} Q(\alpha; \alpha^{w-1})$. We run the EM algorithm using the code developed by [69], which is available in https://github.com/Mogeng/IOHMM.

3.3.5 Predictability of Mobility

Since individual mobility have both regularities and uncertainties [21, 58, 62, 78], it is challenging to evaluate the performance of the prediction models. Previous studies investigated to uncover the relationships between randomness (unforeseeable) or regularity (predictable) of individual mobility behavior mainly using mobile phone data [21] and GPS traces [62]. However, the predictability of ride-hailing service users' mobility behavior is seldom uncovered. Since the characteristics of ride-hailing service users' mobility behavior is different from the movement patters of mobile phone users and regular individuals, it is essential to capture the degree of predictability of ride-hailing service users mobility behavior. Thus, in this study, we have calculated the predictability using the entropy measures based on the historical mobility sequence of each individual. Predictability of mobility sequence can be defined into three ways: random predictability, temporal-uncorrelated predictability, and real predictability. In this study, we use real entropy to calculate the predictability of mobility sequence.

The real entropy (e_{real}) is not only associated with the probability of visited locations in historical data but is also related to the frequency of the order of visited locations. For each individual, the real entropy can be calculated by Equation 20.

$$e_{real} = -\sum_{L_i^{'} \in L_i} p\left(L_i^{'}\right) log_2\left[p\left(L_i^{'}\right)\right] \quad (28)$$

where L_i is the sequence of visited locations of individual *i*; $p\left(L_i^{'}\right)$ is the probability that a specific order of visited places happened in the historical time series.

According to Fano's inequality [75], to uncover the predictability, we calculate the probability Π^{max}, which means the maximum predictability for a specific sequence. For each individual, given the entropy and the number of locations, the predictability of the visited location sequence can be calculated by Equation 21.

$$e_{real} = H(\Pi^{max}) + (1 - \Pi^{max}) log_2(N - 1) \quad (29)$$

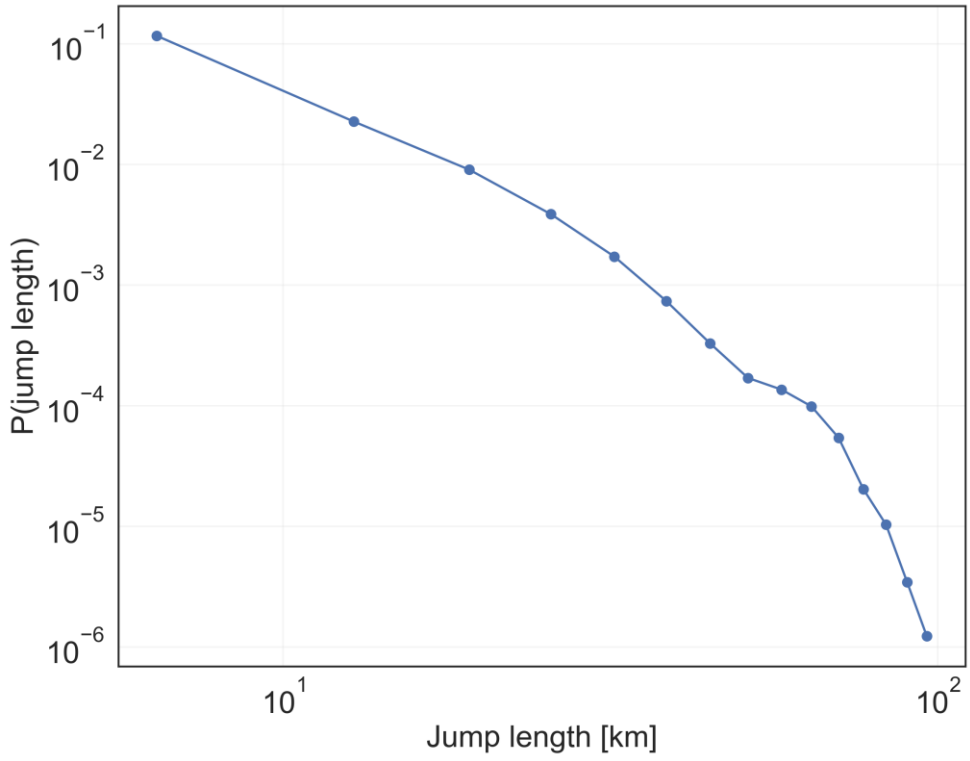
where, e_{real} is real entropy of the individual; $H(\Pi^{max})$ is a function of Π^{max}, which can be calculated by $H(\Pi^{max}) = -\Pi^{max} log_2(\Pi^{max}) - (1 - \Pi^{max}) log_2(1 - \Pi^{max})$; N is the total number of visited locations.

3.4 Empirical Results

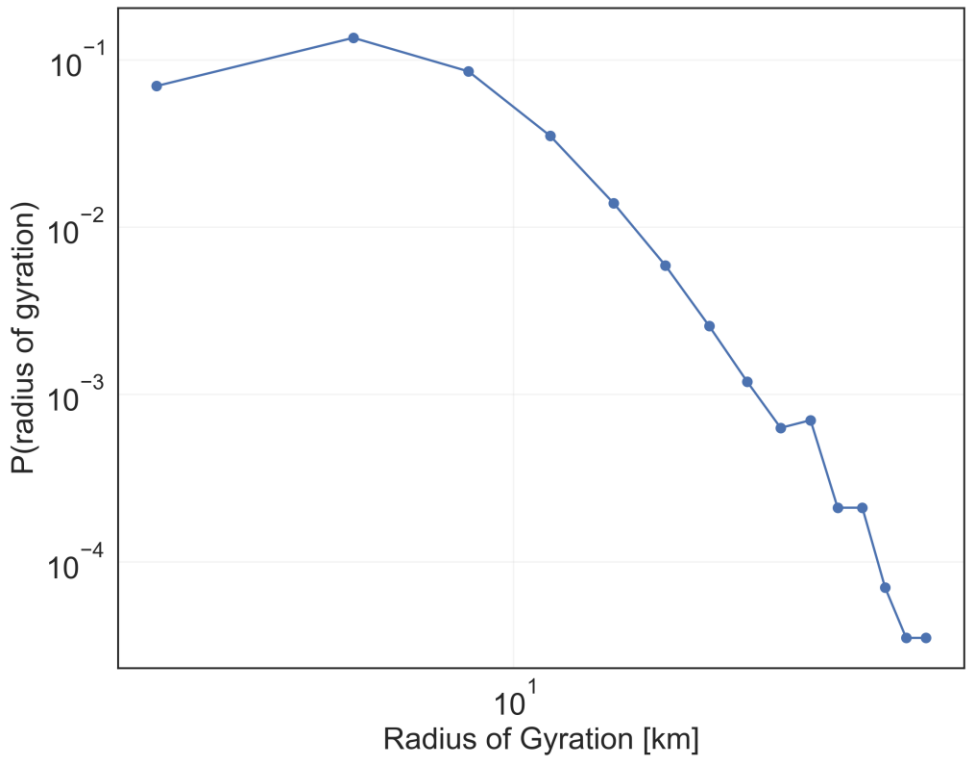
3.4.1 Mobility Patterns

To uncover the spatial-temporal mobility patterns of ride-hailing service users, we have analyzed the distributions of the jump length, radius of gyration and the hourly trip generation volume across the whole user population, shown in **Figure 3.5**. The jump length represents the travel distance of individuals for each trip and the radius of gyration refers to the root mean square distance of all the visited points to the center of mass for each individual's mobility sequence. The jump length and radius of gyration reveal the spatial patterns of user mobility. The hourly trip generation volume shows the ride-hailing service demand in each hour of a day revealing the temporal patterns of user movement.

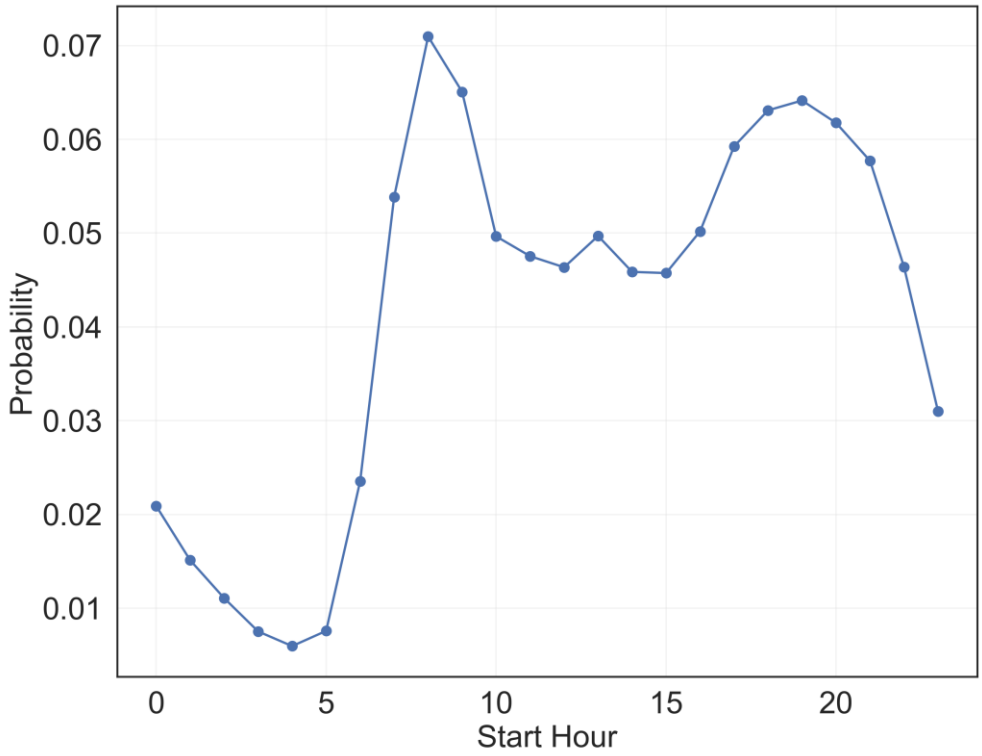
Figure 3.5 (a) indicates that almost all the jump length of ride-hailing service trip is less than 100 km, and the majority of jump length is less than 10 km which shows that individuals seldom travel longer distance when using an on-demand ride-hailing service. Similarly, **Figure 3.5 (b)** shows that most of the individuals have a less than 10 km radius of gyration indicating that users tend to have a small size of activity region when they take the ride-hailing service. For temporal patterns, we find a typical bimodal distribution of hourly trip generation volume including morning peak hours (7 am – 9 am) and afternoon peak hours (6 pm – 8 pm) (see **Figure 3.5 (c)**). It shows a strong temporal regularity of mobility behavior of ride-hailing service users indicating that the hour of the day has a significant influence on mobility patterns.


P(jump length)
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{1}
10^{2}
Jump length [km]

(a)


P(radius of gyration)
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{1}
Radius of Gyration [km]

(b)



(c)

Figure 3.5 Spatio-Temporal patterns of ride-hailing service: (a) the distribution of jump length; (b) the distribution of radius of gyration; (c) the distribution of hourly trip generation

3.4.2 Trip Decision Prediction

In this section, we use real-world on-demand ride-hailing service data to validate the performance of the proposed trip decision model. Due to the heterogeneity in individual mobility behaviors, we train the trip decision prediction model for home-based users, work-based users, commute-based users, and random users, separately. The input features used in this model contain the day of the week (Mon, Tue, Wed, Thu, Fri, and Sat), the number of trips of the individuals (NTS), the number of active days (NAS), the radius of gyration (RGS), the average distance record (ADRS) (distance between a passenger and the vehicle when the on-demand ride-hailing service driver

receives the order), the average travel cost (ACS), the average travel distance (ADS), the average travel time (ATS), the weather (Wea), the air quality (AQ) and the last week active days (LA). To evaluate the proposed trip decision model, we also applied a logistic regression model to predict whether an individual will make a trip as a benchmark model.

Figure 3.6 shows the results of the trip decision model for each group of users. As we can see from the figure, all groups have similar results for the trip decision model with a median accuracy value around 65%. And it also indicates that the proposed trip decision prediction mode works better than the logistic regression model across all the groups. The results demonstrate that the travel purpose of on-demand ride-hailing users does not have significant influence on whether the users will have a trip or not.

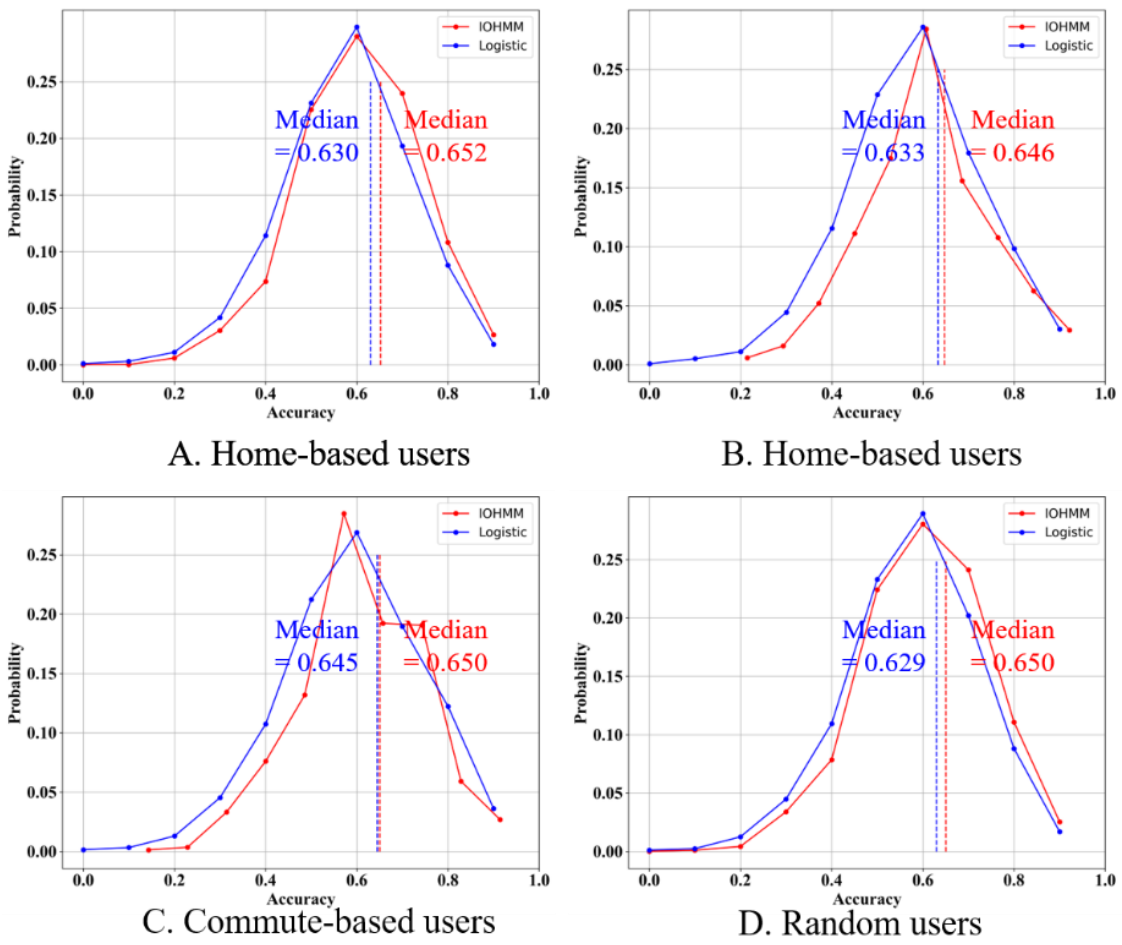


Figure 3.6 Trip decision prediction model results: (a) home-based users; (b) work-based users; (c) commute-based users; (d) random users

In addition, we can also find the patterns of daily trip numbers in the trip decision prediction model. **Table 3.2** presents the emission parameters of the trip decision model. We can find that all the four groups make on average 1.6 to 1.7 trips per day. Since the average number of trips per day is similar across the four clusters, it indicates that the travel purpose does not significantly influence the number of trips made per day. We can find that the number of trips and the number of active days have significant influence on the daily number of trips. Also, the coefficient of the distance shows a negative relationship with the number of trips per day. It

indicates that if an individual has a larger average distance record, she will prefer to make less trips by using a ride hailing service.

Table 3.2 Emission parameters of daily trip number of trip decision prediction model

	Home-based users	Work-based users	Commute-based users	Random users
Constant	1.726	1.64	1.557	1.679
Mon	-0.051	-0.021	-0.016	-0.033
Tue	-0.043	-0.005	-0.008	-0.006
Wed	-0.025	-0.002	0	-0.001
Thur	-0.04	0.001	-0.001	-0.013
Fri	-0.051	0	-0.003	-0.025
Sat	-0.015	-0.016	0.008	-0.004
Number of Trips	5.266	4.588	4.596	5.061
Number of Active Days	-2.536	-2.186	-1.999	-2.516
Radius of Gyration	0.014	-0.008	0.026	0.012
Average Distance Record	-0.809	-0.313	-0.241	-0.062
Average Travel Cost	0.144	0.098	-0.15	0.063
Average Travel Distance	-0.206	-0.18	-0.095	-0.122
Average Travel Time	0.188	0.165	0.245	0.228
Weather	-0.01	0.003	0.009	0
Air Quality	0.002	0.004	0.004	0.002
Last Week Active Days	0.018	0.02	0.015	0.021

3.4.3 Next Origin and Destination Prediction

In this section, we present the results of mobility sequence generation model estimated over the data from the selected 34,311 users extracted from the Didi ride-hailing service platform to generate the mobility sequence of each individual if they have a trip based on trip decision model. In the testing process, we calculated the accuracy for each individual so that we can analyze the distribution of accuracy across all the users in the test dataset to provide a more detailed evaluation of the prediction models. To compare the IOHMM with a benchmark model, we also trained a traditional HMM in the same ways (same training and test datasets). We visualize the results of the origin prediction models in **Figure 3.7** and the destination prediction models in **Figure 3.8**.

In general, for origin prediction, we find that IOHMM outperforms the benchmark model – HMM for every cluster. **Figures 3.7 (a) – 3.7 (d)** demonstrate the distribution of prediction models for home-based users, work-based users, commute-based users, and random users, respectively. It indicates that prediction models of home-based users and work-based users have 52% (50% for HMM) and 50% (46% for HMM) accuracy, respectively. The prediction model for commute-based users achieves the highest accuracy with the median of the accuracy distribution is 71% (51% for HMM) (**Figures 3.7 (c)**).

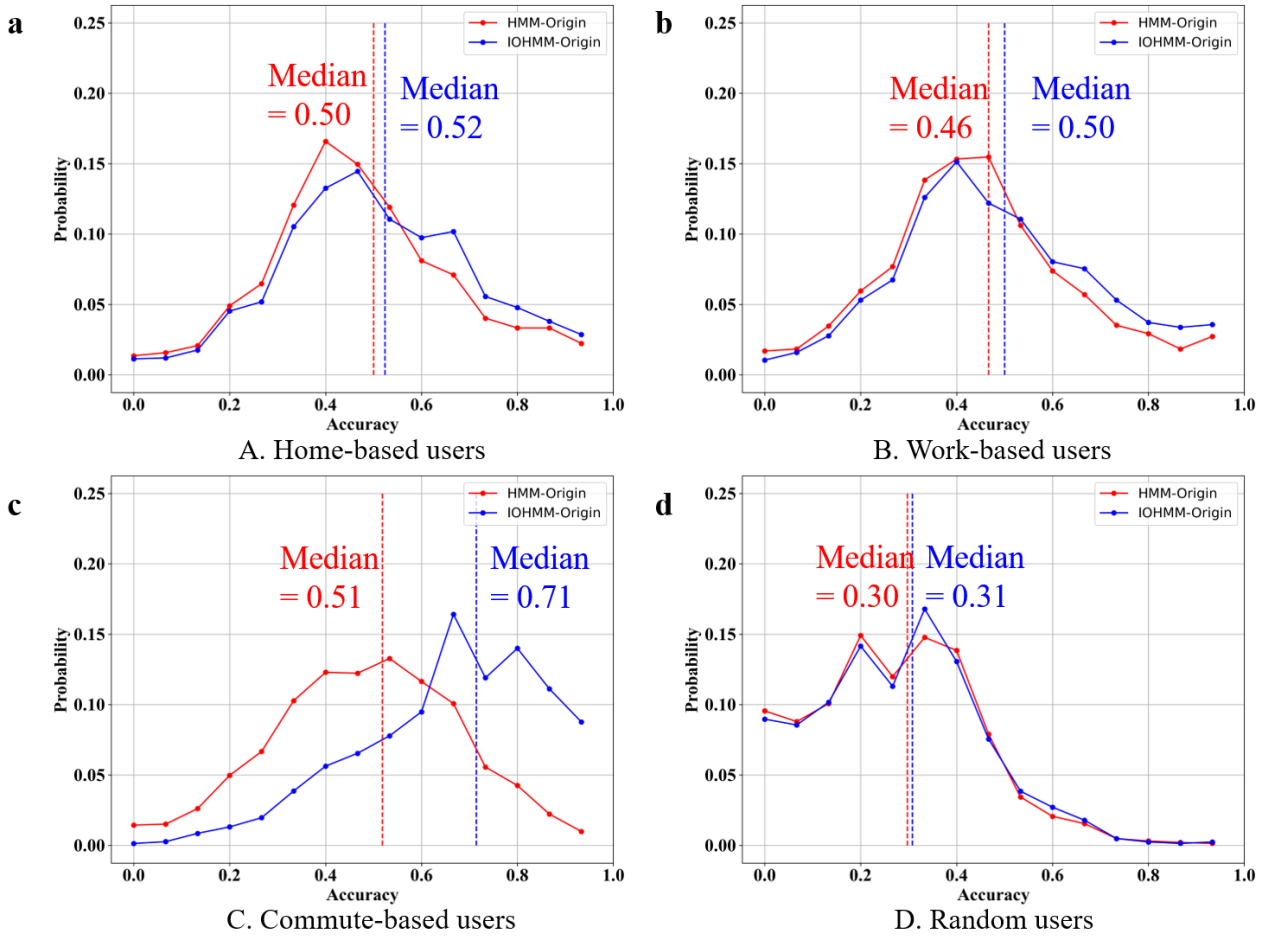


Figure 3.7 Origin prediction model results: (a) home-based users; (b) work-based users; (c) commute-based users; (d) random users

For destination prediction model, it shows the similar results with the origin prediction model in **Figure 3.8**. The commute-based users have the highest accuracy among all the groups. However, the general accuracy of destination prediction is less than the origin prediction. The accuracy of destination prediction model for commute-based users can reach 67% followed by the home-based users (44%) and work-based users (40%). The random users have a low accuracy (33%) which also indicates the randomness present in the mobility behavior of these users. Except for the random users, we found that the IOHMM model outperforms the HMM model over all other user groups.

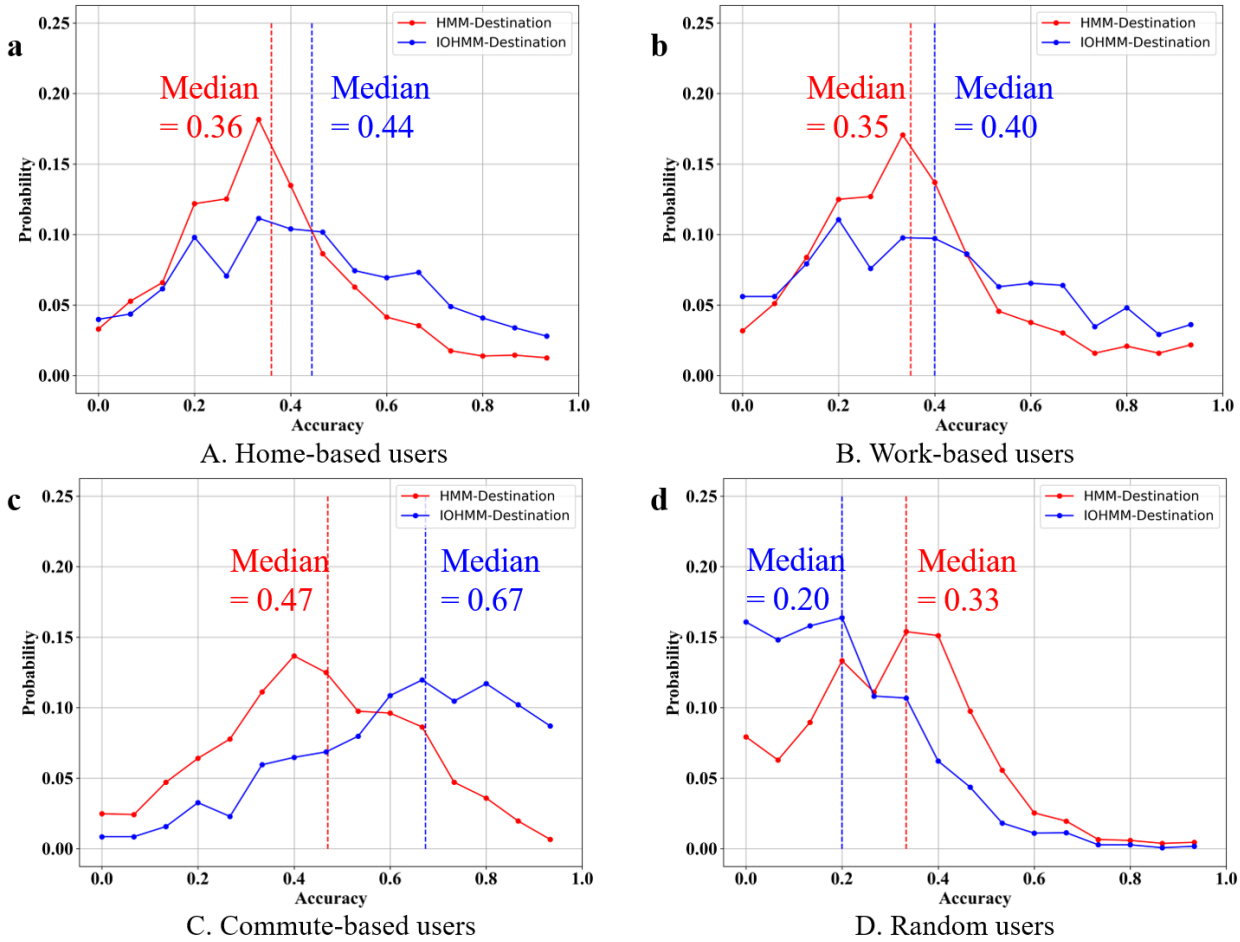


Figure 3.8 Destination prediction model results: (a) home-based users; (b) work-based users; (c) commute-based users; (d) random users

We have compared our results with previous studies that developed individual travel behavior prediction models. Using transit smart card data, Zhao et al. [25] developed an n-gram model to predict the trip decision, origin and destination of the users. They reported accuracy values of 80% for the trip decision prediction, 66.7% for the origin prediction, and 46.7% for the destination prediction. Since most smart card users are likely to travel for a commuting purpose, their mobility behavior may have more regularities than ride-hailing service users. Our results show that for commuting-based users our model can also achieve accuracy levels similar to that for smart card users. In another study, Lv et al. [31] used data from 3000 mobile phone users to develop an HMM for the prediction of the next place to visit at an individual level. The results show that for individuals with a more regular lifestyle, such as those classified as "family persons", an accuracy value between 60% and 70% can be achieved in most of the time, while for individuals with less regular lifestyles only an accuracy value between 20% and 50% can be achieved. It is worth noting that the prediction accuracy for individuals with regular travel patterns is relatively high, which is similar to our findings that for commuting-based users higher accuracy levels can be achieved. Individual mobility patterns captured by transit smart card transactions or mobile phone records are likely to have more regularities; hence, models developed over such data are likely to have better accuracy levels. Considering the randomness in the mobility behavior of ride-hailing service users, we believe that the models presented in this paper are producing promising results.

3.4.4 Temporal Patterns of Model Accuracy

In addition, we investigated the relationship between temporal features and model accuracy. To calculate the model accuracy over different time periods, we extracted both the ground truth and the predicted values given by the models for each specific period (indicated by time of the day

and day of the week). To visualize the correlation between the temporal features (time of the day and day of the week) and model accuracy, we plot the accuracy distribution of mobility sequence generation model (origin and destination prediction) across the user groups over different temporal variables, which can be seen in **Figures 3.9(a) – 3.9(d)**. We found that in the morning and afternoon peak hours, the model accuracy will be higher. For example, the origin prediction can achieve more than 90% accuracy in the morning peak hours for home-based users and commute-based users. **Figures 3.9(c)** and **3.9(d)** also show that the day of week can also influence model accuracy. For all home-based, work-based, and commute-based users, the accuracies for weekends are lower than the accuracies obtained from the same users over weekdays.

The findings that both time of the day and day of the week have significant influence on model accuracy show that the mobility sequence generation model can capture the temporal patterns of the context information. On peak hours and weekdays, since individuals always have predictable mobility behavior (commuting), the model accuracy should be higher than other time periods.

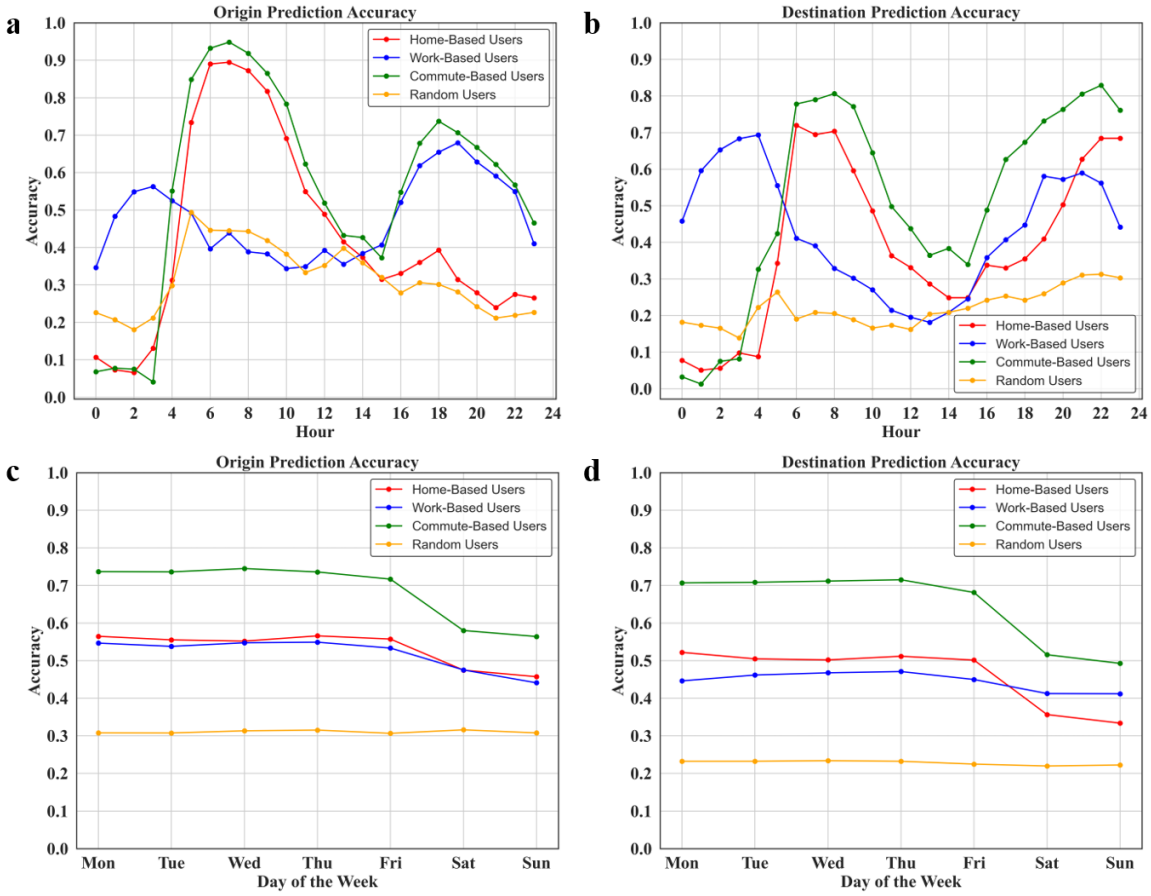


Figure 3.9 Model accuracy for different temporal features for each cluster

3.4.5 Predictability vs. Model Accuracy

Since ride-hailing service users are likely to have uncertainties in their travel patterns, to better evaluate the performance of the mobility sequence generation model, we calculated the predictability of individual mobility sequence based on Fano's inequality [21, 75]. The patterns of the predictability (see **Figure 3.10 (a))** are consistent with entropy patterns observed in previous studies that the real predictability is higher than the time-uncorrelated predictability [21]. It reports that the distribution of real predictability peaks at around 60% with a lower bound at about 40% across all the individuals. The results indicate that a model can predict the origin of the ride-hailing service trips for 60% of the time. As such, predictability values can be considered as the limits of accuracy levels for mobility prediction models.

To evaluate the relationship between predictability and model prediction accuracy, we also investigated the correlation between the calculated predictability with model accuracy for each group. **Figure 3.10 (b)** shows that the prediction accuracies are correlated with the real predictability of individual mobility sequence. For an increase in the real predictability values, the prediction models' accuracy of all the four user groups will also increase. We can find that the commute users (green points in **Figure 3.10 (b)**) have the highest predictability (average = 73%) and the random users (orange points) have the lowest predictability (average = 58%). It shows the model accuracy levels for most of the commuter users, home-based users and work-based users are close to the corresponding predictability values.

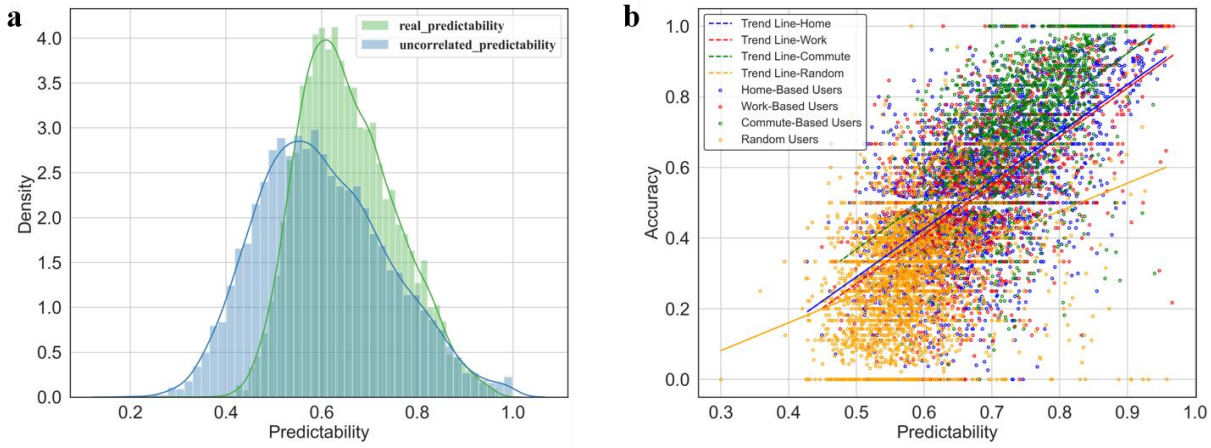


Figure 3.10 The results of mobility behavior predictability

3.5 Conclusions

In this research, based on the historical mobility sequence, we applied an innovative mobility prediction model – a supervised learning-based multi-layer hidden Markov model to predict trip decision and the next ride-hailing service trip at an individual level—using massive ride-hailing service data considering the heterogeneity of travel purpose. To the best of our knowledge, this is the first study to develop a model to predict the mobility behavior of on-demand ride-hailing service users at an individual level. Considering the heterogeneous mobility patterns of ride-

hailing service users, we divided individuals into four groups – home-based users, work-based users, commute-based users, and random users. We train the proposed model separately for each group. From the results, we can find the trip decision model works better than the logistic regression model which can achieve an accuracy of around 65%. The emission parameters also capture the patterns of the daily trip number of the individuals. Besides, it shows that the mobility generation model can work well for home-based users, work-based users and commute-based users for origin and destination predictions. We also investigated whether the temporal variation would have an influence on the accuracy of the proposed model for each group. It indicates that the accuracy of prediction models is higher in peak hours and weekdays for home-based users, work-based users, and commute-based users. To validate the performance of the proposed individual mobility prediction model, we also checked the entropy and predictability of the mobility sequence for each individual. The results show that the distribution of predictability peaks at around 60%. In terms of the correlation between the predictability and prediction model accuracy, it reveals that the accuracy levels of model predictions for each user group are proportional to the corresponding predictability values of their mobility sequence. This is an important discovery since it means that the predictability method can be used for model improvement and evaluation in the future.

This study has some limitations such as: (i) since the visited places with a higher rank are challenging to predict, we cluster all visited places with a rank higher than 10 as low-frequent visited locations. As such, the model cannot forecast any place with the high rank as the next visited place. To solve this problem, future work should focus on identifying activity types of each ride-hailing trips (such as home, work, foods or shopping trip); (ii) due to the lack of available data on individual characteristics (e.g., age, gender or job), the proposed model cannot

have higher performance since it cannot capture individual characteristics contributing to the heterogeneity. In future research, with more individual-level socio-demographic data, the proposed model can be improved. However, such information about the ride-hailing service users will be hard to obtain due to privacy concerns.

CHAPTER 4: ASSESSING THE IMPACTS OF A REAL-TIME RIDESHAING SYSTEM USING AN AGENT-BASED SIMULATION MODEL

4.1 Introduction

[2] Nowadays, increasing growth of population and private vehicles in most cities cause severe congestion problems and associated environmental pollution [15]. To mitigate the congestion, fuel consumption and pollution (e.g., greenhouse gases (GHG) emissions), one of the major solutions is to improve the efficiency of transportation modes using emerging trends such as car sharing and ridesharing. Recently, with the wide adoption of online ride-hailing services such as Uber, Lyft, and Didi, the number of options to move individuals has increased. Users can now easily use a smartphone app to request a ride-hailing service. This provides a new opportunity to bring individual passengers together with similar routes and time schedules, to decrease the number of service vehicles and to reduce the negative impacts of transportation systems.

With on-demand ride-hailing service platform, a user can request a service vehicle with the origin and destination of her trip. And then the platform will assign one of the nearest available service vehicles to serve the request. However, currently, most of the ride-hailing service vehicles only serve for one person or group, which lowers the efficiency of the service vehicle. To improve the efficiency of the ride-hailing service, dynamic ride-sharing strategy can be utilized.

The ridesharing strategy can support more than one person or group in one service vehicle, which can decrease the fleet size, and save costs to the users. It has also been claimed that dynamic ridesharing can mitigate traffic congestion in high density areas [16]. However,

[2] Zhang, J., Hasan, S. and Yan, X., 2023. Assessing the Impacts of a Real-time Ridesharing System using an Agent-based Simulation Model. In Transportation Research Board 102nd Annual Meeting. (accepted)

since the ridesharing may have to detour to meet all the users in the vehicle, the travel mileage of the users will increase. Thus, assessing the impact of ridesharing is important for the future policy making.

In this research, we focus on assessing the travel and environmental impacts of ridesharing strategies. We built a real-time ridesharing system using an agent-based simulation model to extract the trajectories of passengers and vehicles which can provide valuable information. To reflect more about the real-world situation, we use the real demand data extracted from Didi ride-hailing service platform and traffic condition data from Google Map API. In addition, we proposed a heuristic matching algorithm (vehicle - passenger) for both ridesharing and non-ridesharing system in real-time so that all passenger requests can be served.

The contribution of this paper can be summarized as follows:

- We built an agent-based model to simulate the operations and impacts of both ridesharing and non-ridesharing systems with a real-time matching algorithm.
- We use real world on-demand ride-hailing service data and real traffic status data to test the operations of both ridesharing and non-ridesharing systems.

4.2 Literature Review

Since the high density of population and vehicle ownership occur in many large cities, such as New York, London, Beijing, and so on, congestion has been one of the major concerns for transportation system management. Thus, how to improve the efficiency of different transportation modes has been an important research area. In addition, with the wide availability and adoption of ride hailing services, policymakers are concerned with how to improve the efficiency of these services and reduce their negative impacts on city traffic. A potential solution is to offer a ride-sharing option within these ride-hailing services.

Initially, researchers focused on ridesharing with fixed stations [82, 83] or pre-arranged trips [84, 85]. For the fixed stations, they preset the taxi stations and the passenger should arrive at the station first to take a taxi. For the pre-arranged trips, individuals with similar characteristics (ex. start location, age, gender, et al.) will negotiate first to decide whether they will make a car pooling. However, with the development of internet technology, the on-demand ride-hailing service has become more popular in individuals travel mode. It requires the service vehicle to satisfy the passengers' requests in real time. Thus, the fixed stations and pre-arranged method cannot meet the requirements of the on-demand ride-hailing service users, which cannot evaluate the impacts of the ridesharing or carpooling for these users.

In recent years, researchers have focused on analyzing the influence of ridesharing or carpooling strategies on on-demand ride-hailing service users [17, 86-88]. It has been stated that the ridesharing strategy will mitigate traffic congestion [16], reduce greenhouse gas emissions [17], save travel cost [89, 90], and decrease vehicle-miles traveled (VMT) [91, 92]. Alisoltani et al. [16] proposed a complete framework to operate the ridesharing system using a dynamic trip-based macroscopic simulation in two cities with different trip density. The results show that the ridesharing strategy can significantly mitigate the congestion problem in high demand density areas. However, in cities with low shareability, ridesharing will cause extra travel mileage. Fagnant and Kockelman [93] analyzed the impacts of shared autonomous vehicle (SAV) using agent– and network–based simulation with a dynamic ridesharing strategy. The study suggests that the application of SAV can improve the service quality and save travel costs of passengers. In addition to the impacts on traffic conditions, environment and economy, ridesharing can also affect urban parking demand. Zhang et al. [94] used an agent-based simulation model to evaluate the potential influence of ridesharing on urban parking demand under different scenarios. It

shows that up to 90% of urban parking demand can be reduced if enough SAVs are deployed in the system.

Based on the aforementioned studies, agent-based simulation model (ABM) has been widely adopted in optimizing the ridesharing system or evaluating the impacts of the ridesharing strategy [94-97]. The agent-based model contains a collection of agents which can make decisions automatically based on preset rules [98]. The agent-based model is a powerful tool to simulate the dynamic travel behavior in the real world, even if the model is simple [99]. In the ridesharing system, the service vehicles and passengers are always regarded as the agents. With the operation of the agent-based simulation model, the motion patterns of agents can be recorded which can provide valuable information for analyzing the performance of the system. Besides, the movement of agents in the agent-based simulation model relies on a set of rules [98]. However, most of the previous agent-based models for ridesharing systems made simplified assumptions, regarding traffic demand, travel distance and travel time, without considering real-world demand and traffic conditions [87]. In Fagnant and Kockelman [86], an agent-based simulation model is operated with generated trips, trip distance and fixed travel speed. Since the study area and the trips in the system are based on assumptions, it is difficult to reflect the real-world situation. To improve the reality of the simulation, Liu et al. [100] generate the trip with an activity-based simulation model in the City of Austin. This study investigated the influence of different fare levels of SAV on the operation of ridesharing system. The results indicate that with a higher fare, the percentage of SAV demand in total trips will decrease.

One of the most essential challenges in a ride-sharing system is how to match the user and service vehicles efficiently. Since the vehicle will take more than one rider group, it will lower the efficiency of the service if the shared rider groups do not have enough overlapping of

their routes. In recent years, researchers have proposed several ridesharing matching algorithms [87]. Lokhandwala and Cai [87] proposed an algorithm using two rectangular bounding boxes to match the users and service vehicles. The first bounding box contains all the pick-up and drop off points of current trip chain of the service vehicles. A new origin-destination location should be in the first bounding box to share the vehicle. Otherwise, the second bounding box is formed with both the new and previous pick-up and drop off points. If the second bounding box covers the first bounding box, then it allows the new requests to be shared with the previous trip chain. However, the proposed algorithm can only group the riders with similar directions. If the bounding box of trip chain is large, the time of detour for ridesharing will be long, which will make it inefficient of the service. In addition, there is another type of matching algorithm which requires the users to request the service vehicle in advance and then the system will make the schedule based on the requests [100]. However, this matching strategy will limit the ability of ridesharing services. Individuals prefer to be served immediately rather than request a service vehicle many hours in advance. Thus, a real time matching algorithm will be more useful in ridesharing systems.

Nowadays, on-demand ride-hailing services play an essential role in urban transportation research providing large-scale valuable data for passengers' mobility behaviors with a potential in developing ridesharing systems for real world scenarios. From above mentioned literature, there is limited research focusing on real-time vehicle-passenger matching methods. And the simulation models in most of the previous studies are based on assumptions which may not reflect real-world situations. In this research, we have developed an agent-based model to evaluate the potential impacts of ridesharing system with a real-time vehicle-passenger matching

algorithm using real-world trip demand data from the Didi platform and traffic conditions data from Google map API.

4.3 Data and Methods

4.3.1 Agent-based Model

In this study, we build an agent-based model to simulate the process of all the ridesharing services in the system. In the agent-based model, we have two types of agents – passengers and vehicles. We can collect all the parameters of the two types of agents which can provide the data for statistical analysis both at a system-level and an individual-level. To compare with the ridesharing, we also built a basic agent-based model to simulate the scenario without a ridesharing option.

We assume that the system would operate in a real-world city, Beijing, as we have the ride hailing demand data available for the city. To save the cost of computation, we selected one of the busiest regions in Beijing, which can be seen in **Figure 4.1**. The study area is 5 * 5 km^2 rectangular. To generate and attract trips, we divided the whole area into 40 * 40 grids. And we also use the grid to update the position of passengers and service vehicles.

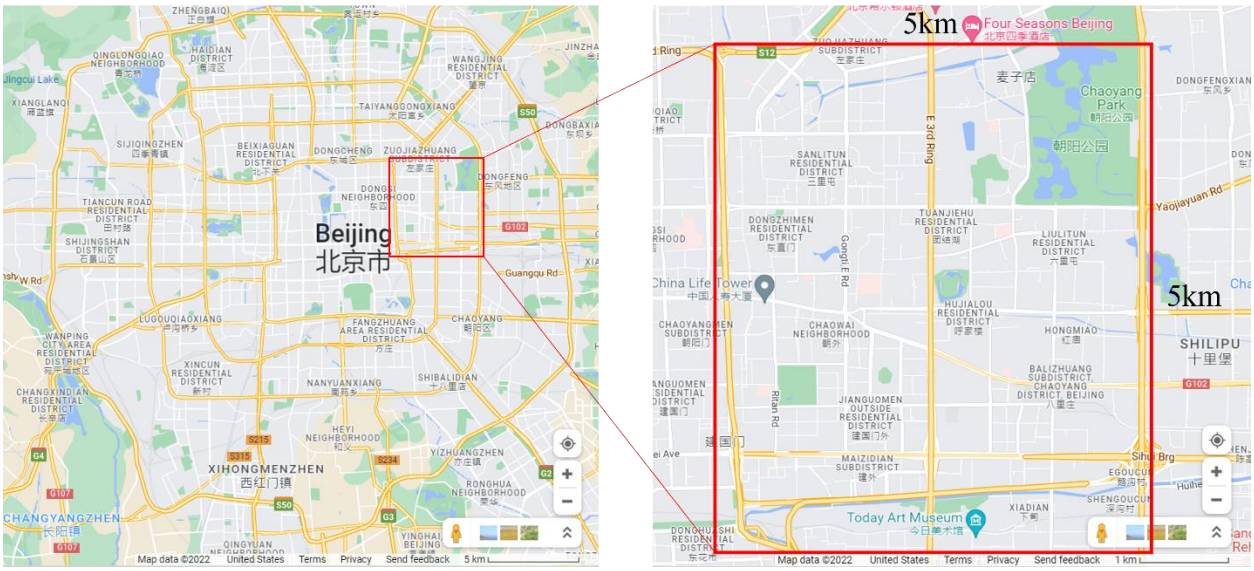


Figure 4.1 Study area

We initialize the simulation with the earliest timestamp of the dataset and update the system (the position of the service vehicles and passengers) every 10 seconds. To reflect the real traffic status, we extracted the travel distance and travel time of all the OD pair from Google Map API (https://developers.google.com/maps/) based on the timestamp.

However, the Google Map API can only provide the travel time and travel distance with the best route for current time and a time in the future. We collected the travel time and travel distance matrix for four periods – morning peak hour (7am-10am), afternoon (10am - 7pm), afternoon peak hour (7pm – 9pm), and evening (9pm – 7 am). Then, we matched the departure time of each trip with these four periods to obtain the real travel time and travel distance.

Once there is a request from a passenger, the system will immediately check the nearest available service vehicle based on our matching algorithm. We present the details of the matching algorithm below. For the scenario without ridesharing, the capacity of the service vehicle is set as 1 and for the scenario with ridesharing the capacity of service vehicle is set as 2.

4.3.2 Matching Algorithm - Base Scenario

For the scenario without ridesharing, if there is a request from a passenger, we will simply find the nearest empty service vehicle and then update the position of the vehicle, until the vehicle drops off the passenger.

We match the passengers and service vehicles using the following steps:

Step 1: if there is a request, find the nearest empty service vehicle based on the travel distance extracted from Google Map API. After the match is scheduled, change the service vehicle status to 0.5 (matched but empty service vehicle).

Step 2: Set the origin of the matched but empty vehicle as the current position of the vehicle and the destination as the origin of the matched passenger. Then, update the position of the service vehicle until the vehicle picks up the passenger (arrive at the destination of the service vehicle), and change the service vehicle status as 1 (occupied service vehicle).

Step 3: After the service vehicle drops off the passenger, set the service vehicle status as 0 (empty service vehicle).

Step 4: Run step 1-3 continuously until all the requests are satisfied.

4.3.3 Matching Algorithm - Ridesharing Scenario

For the scenario with ridesharing, when there is a request from the passenger, we will not only consider the empty service vehicle, but also find the vehicle which has not reached its capacity. However, if the vehicle already has 1 passenger, then we also need to consider the service quality (total travel time and comfort level) of the current passenger. Thus, in this study, we set the capacity of the shared service vehicle as 2. Besides, we also calculate the detour time to verify whether the service vehicle is available based on a preset threshold.

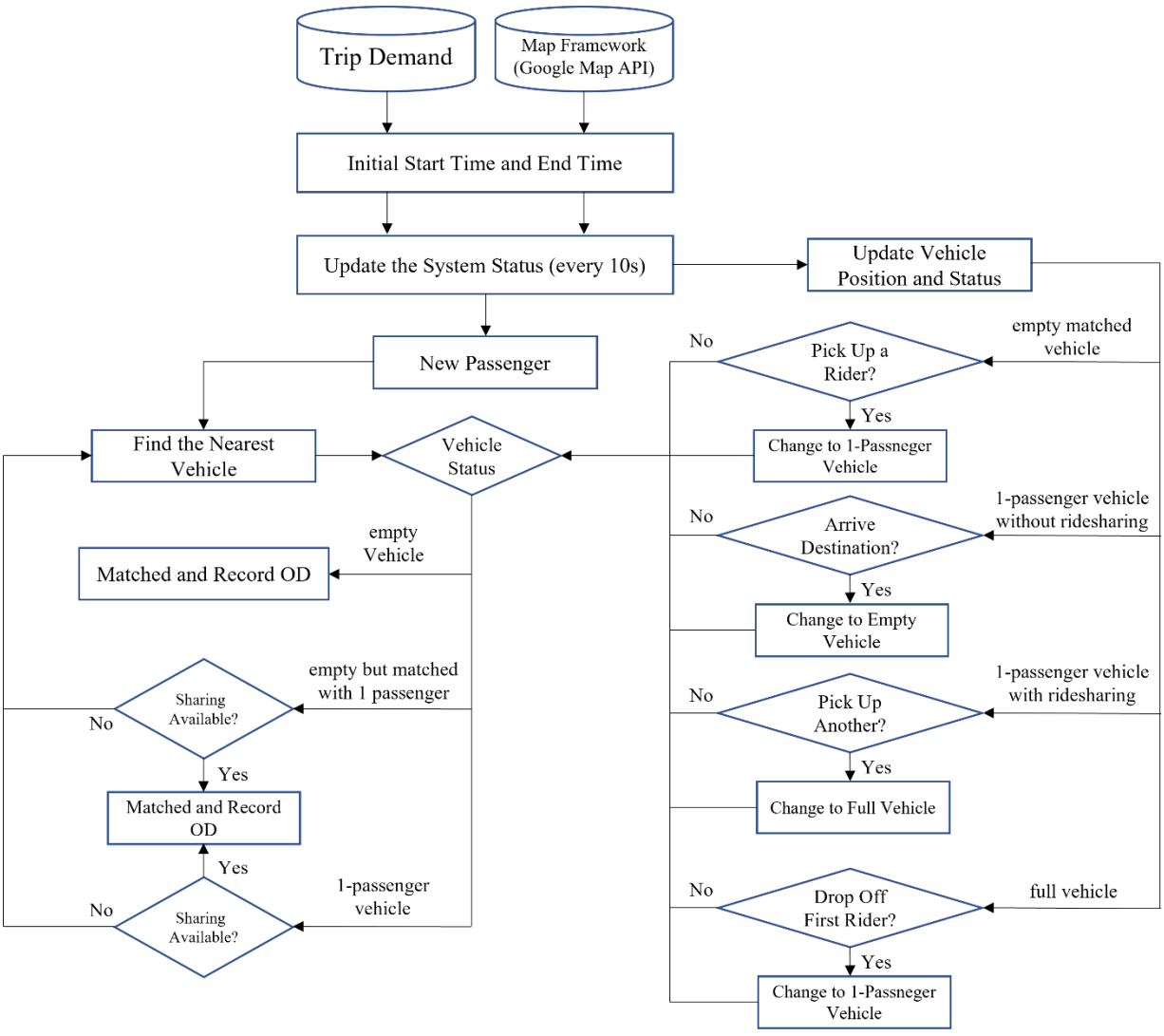


Figure 4.2 Ridesharing simulation modeling framework

The ridesharing simulation modeling framework can be seen in **Figure 4.2**. We match the passengers and service vehicles using the following steps:

Step1: if there is a request, then find the nearest available service vehicle based on the travel distance extracted from Google Map API. After the match is scheduled, change the service vehicle status to 0.5 (matched service vehicle which is not full). If the matched service vehicle is empty, then the service vehicle is available for the request. Otherwise, we need to verify whether

the service vehicle is available based on the following rules. After we find the nearest available service vehicle, we change the status of the service vehicle to 0.5 and record the best route.

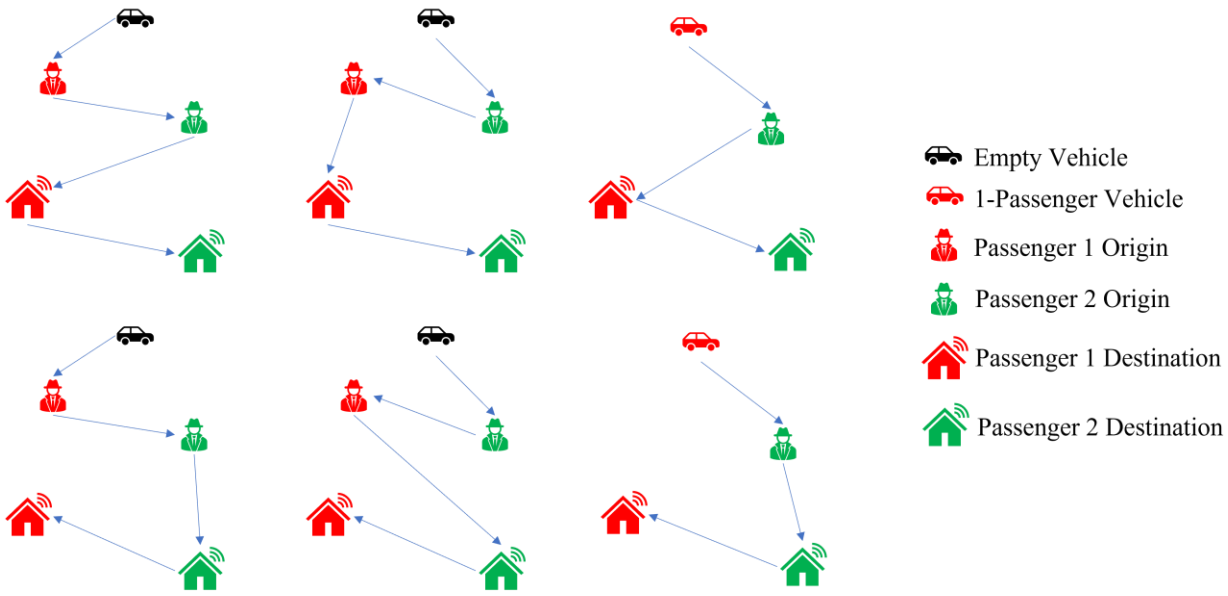
To check whether a service vehicle is available for ridesharing, first, we should calculate all the potential routes for the paired passengers considering the vehicle position, the origin and destination of the two passengers. There are two types of service vehicles when matching with two users – empty service vehicle and 1-passenger service vehicle, which can be seen as **Figure 4.3**.

For empty service vehicles, based on the algorithm, there should be one passenger matched earlier and the matched service vehicle does not pick up the passenger at the timestamp when the second passenger requests. In this situation, there are 4 potential routes for the two passengers, shown in **Figure 4.3 (A)**. The 4 potential routes are separately:

- empty service vehicle – passenger 1 origin – passenger 2 origin – passenger 1 destination – passenger 2 destination
- empty service vehicle – passenger 1 origin – passenger 2 origin – passenger 2 destination – passenger 1 destination
- empty service vehicle – passenger 2 origin – passenger 1 origin – passenger 1 destination – passenger 2 destination
- empty service vehicle – passenger 2 origin – passenger 1 origin – passenger 2 destination – passenger 1 destination

For 1-passenger service vehicle, based on the algorithm, there should be one passenger matched earlier and the matched service vehicle has picked up the passenger at the timestamp when the second passenger requests. In this situation, there are 2 potential routes for the two passengers, shown in **Figure 4.3 (B)**. The 2 potential routes are separately:

- 1-passenger service vehicle – passenger 2 origin – passenger 1 destination – passenger 2 destination
- 1-passenger service vehicle – passenger 2 origin – passenger 2 destination – passenger 1 destination



(A) Ridesharing with Empty Vehicle (B)Ridesharing with 1-Passenger Vehicle

Figure 4.3 Potential routes for ridesharing strategy

We calculate the total travel time for each potential route and the associated detour time for each matched passenger. To satisfy the comfort level of the passengers, we also set up rules to avoid that the passenger will not have excessive detour distance caused by the ridesharing.

The rules are:

- If the matched service vehicle is empty, the waiting time for each passenger cannot exceed the maximum waiting time.
- If the matched service vehicle is empty, the detour time for each passenger should be less than 40% of the travel time that they would have spent without ridesharing.

- If the matched service vehicle has 1 passenger, the waiting time for the second passenger cannot exceed the preset maximum waiting time.
- If the matched service vehicle has 1 passenger, the remaining travel time for the first passenger should be less than 40% of the initial remaining travel time that he/she will spend without matching the second passenger.

Step 2: For the service vehicles without ridesharing, set the origin of the service vehicle as the current position and the destination as the origin of the matched passenger. Then, update the position of the service vehicle until the vehicle picks up the passenger (arrive at the destination of the service vehicle), and change the service vehicle status as 1 (1-passenger service vehicle).

For the service vehicles with ridesharing, we set the status of empty vehicles and 1-passenger vehicles separately. For the empty vehicles, the service vehicle will go to the first passenger based on the recorded best route, until the vehicle picks up the first passenger and then changes the service vehicle status as 1 (1-passenger service vehicle). For the 1-passenger vehicles, they will directly go the *step 3*.

Step 3: For all the 1 – passenger vehicles, there are two types – vehicles without ridesharing and vehicles with ridesharing. If the 1-passenger vehicles do not have the second matched passenger, then it will go to the destination of the current passenger and change the status to 0 (empty vehicle). If the vehicles have a second matched passenger, then the vehicle will go to pick up the second passenger and change the status of the vehicle to 1.5 (1-passenger to pick up another).

Step4: After the second user is picked up, set the status of the vehicle as 2 (full vehicle).

Step5: For all the full vehicles, based on the recorded best route, they will go to the first destination. Once the vehicle arrives at the first destination, the vehicle will drop off one of the passengers and the service vehicle will change to 1-passenger vehicle. Then, we will change the status of the vehicle as 1 (1-passenger service vehicle). Then we will do step *3* for the vehicle.

Step 6: Run *step 1-5* continuously until all the requests are satisfied.

4.3.4 Empty Service Vehicle Relocation

Due to the unbalanced spatial distribution of trip requests in the system, the empty service vehicle in low-demand areas will be inefficient if we do not relocate these vehicles, because the vehicles may stay a long time for the next user. Thus, in this paper, we will relocate the empty service vehicles from the low-demand areas to high-demand areas. Since the raw grid of the system is too small to calculate the demand, we divided the whole study area into 8*8 large grids. We will consider both the supply and demand of these large grids to decide how to relocate the empty service vehicles.

In the system, for each 5-min, we will relocate all the empty service vehicle by following rules, shown as **Figure 4.4**:

Step1: we calculate the aggregated trip requests number D^i for each neighboring large grid and current large grid i (**Figure 4.4** testing large grid) of the current large grid of the empty service vehicle.

Step2: we calculate the number of empty service vehicles and non-full vehicles for each neighboring large grid and current large grid i as S^i. Then for each neighboring large grid and the current large grid i, we calculate the difference between S^i and D^i as T^i. We define the target relocation large grid as the gird with the minimum T^i.

Step3: for the empty service vehicle, we set up the origin as the current location and the destination as the center of the target relocation large grid. Then the system will update the empty vehicle's position until the vehicle arrives at the destination or matches with a new user.

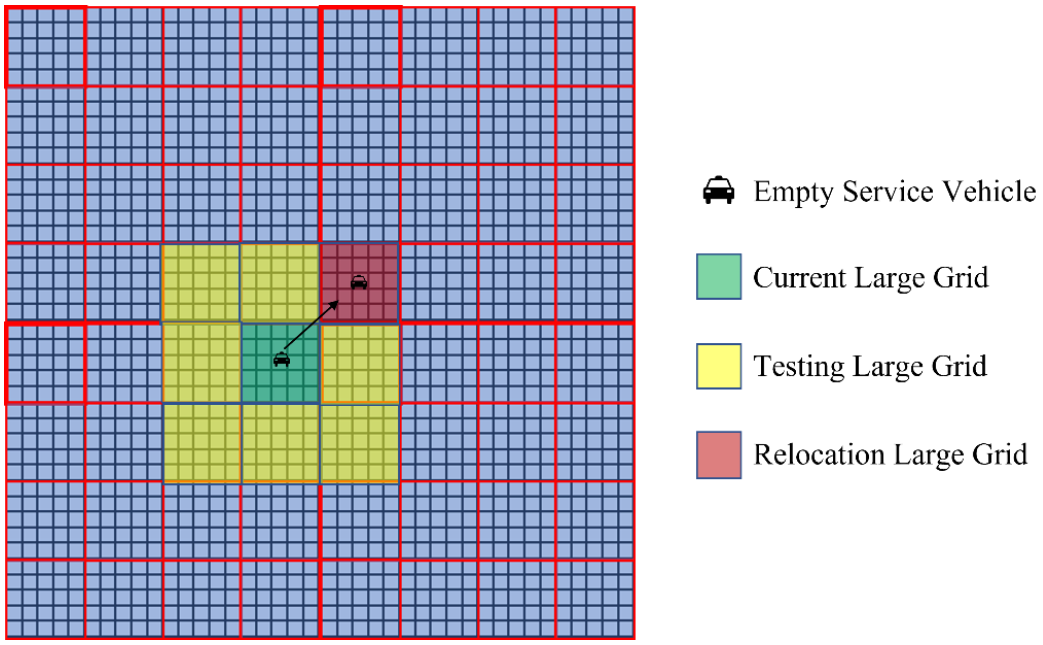


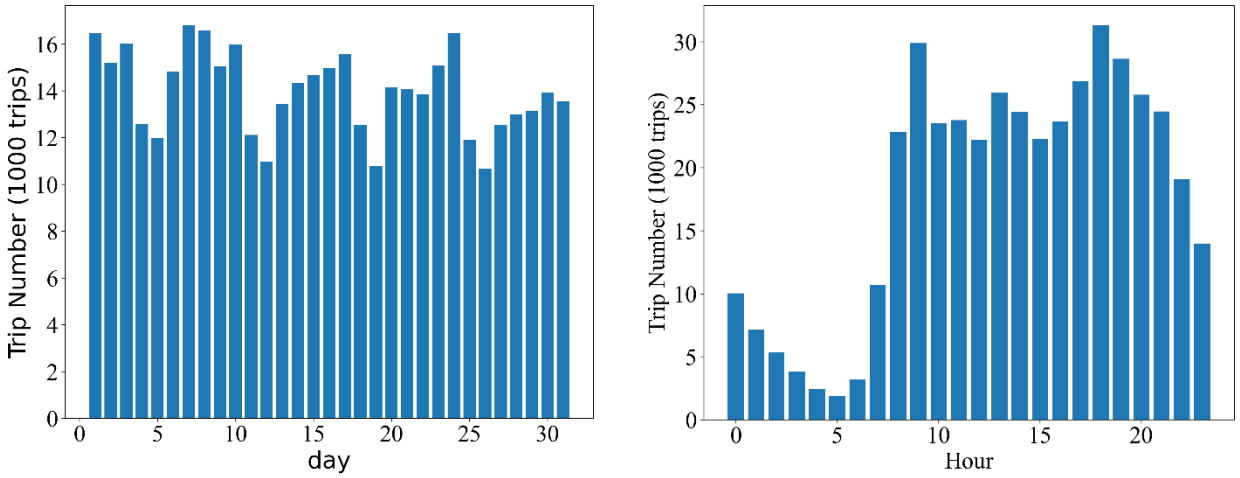
Figure 4.4 Empty service vehicle relocation strategy

4.4 Simulation Results

In this study, we collected one month trip data (March 2017) from Didi ride-hailing service platform for the simulation. Since we selected a subarea in Beijing, we extracted all the trips from which both the origin and destination are in the selected subarea. The data contains 433,242 trips (average 14,441 trips per day), including the request timestamp, the origin and destination of the passengers, and the travel distance. We plotted the temporal distribution of trip requests in two scales – daily and hourly, shown as **Figure 4.5**. We aggregated the demand of requests in hourly level across the whole month data. For example, in Figure **4.5(B)**, the requests number shown in hour 0 is the aggregated trip number of the whole month at time 0am. The demand of the data shows typical weekly periodicity for the passenger's travel behavior. From **Figure**

4.5(A), most of the on-demand service users tend to travel more frequently on weekdays compared to weekends. For each day, it shows most of the passengers start their movements at 8am and ends at 11pm from **Figure 4.5(B)**. The travel behavior of on-demand service passengers also has morning peak hour (9am) and afternoon peak hour (6pm).

We generate the demand in the simulation system based on the real demand data from the on-demand service platform. Although we operate our simulation model in the Beijing area, the proposed approach can also be applied in other regions. To run the simulation, we update the system with the frequency of 6 per minute. For each time window (10s), we update the position of the vehicles and match the latest requests of passengers.



(A) (B)

Figure 4.5 Daily (A) and Hourly (B) distribution of trip number

4.4.1 Fleet Size vs. Maximum Waiting Time

To initialize the simulation, we need to set up the fleet size of the system. The fleet size of service vehicles is related to the waiting time of the passengers. If passengers can wait longer for a service vehicle, it will lower the fleet size of service vehicles in the system. Thus, we first analyze the relationship between the fleet size and the preset maximum waiting time of the

passengers with the following steps. We take one day demand in the simulation including 16,356 total trips.

Step 1: Set a maximum waiting time T_w.

Step 2: Run the simulation with the matching algorithm. If there is a request, check whether there is an available vehicle around the passenger. Then, generate a new service vehicle at the position of the request if no available service vehicle for the passenger in the system.

Step 3: Run the simulation for one day until all the requests are satisfied by the service vehicle. Record all the service vehicles and the final number of service vehicles will be fleet size given the maximum waiting time.

Following the steps, we recorded the maximum number of vehicles as the fleet size. We set up different maximum waiting time in the simulation system – from 2 min to 15 min. The fleet sizes for different maximum waiting time are summarized as **Table 4.1**. From the result, we can find that ridesharing strategy can significantly decrease the fleet size regarding the same maximum waiting time. For the scenario without ridesharing, the fleet size starts from 2093 with a 2-min maximum waiting time. With the increase of the maximum waiting time, the fleet size decreased to 428 with a 15-min maximum waiting time. The pattern is the same for the scenario with ridesharing. However, the fleet size starts from 1406 with a 2-min maximum waiting time for the scenario with ridesharing, which is 33% lower compared to the fleet size (2 min maximum waiting time) without ridesharing. The fleet size reduces as the maximum waiting time increases for the ridesharing scenario up to 349 with a 15-min maximum waiting time. Besides, compared to the fleet size without ridesharing, the fleet size with ridesharing decreases more rapidly until the maximum waiting time reaches 7 min. When the maximum waiting time is 7 min, the fleet size decreases 53% for the ridesharing scenario compared to the scenario without

ridesharing. If the maximum waiting time is higher, the difference in fleet size between the two scenarios will be lower.

Table 4.1 Relationship between the maximum waiting time and fleet size

	Maximum Waiting Time (min)													
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Non-Ridesharing	2093	1584	1430	1177	1009	941	806	705	621	590	507	468	452	428
Ridesharing	1406	940	817	615	498	445	400	380	357	353	356	354	346	349
Decrease Rate	33%	41%	43%	48%	51%	53%	50%	46%	43%	40%	30%	24%	23%	18%

In addition, we also compare the unoccupancy rate between the ridesharing scenario and the scenario without ridesharing. The unoccupancy rate is defined as the ratio of the number of unoccupied service vehicles over the total number of service vehicles. The unoccupancy rate can reflect the efficiency of the service system, since with more unoccupied service vehicles, the efficiency of the system will be reduced. From the results (**Table 4.2**), we can find that the unoccupancy rate will decrease when the maximum waiting time is higher, which means that the efficiency of the service vehicle will increase if passengers can wait more time. It indicates that with ridesharing, the unoccupancy rate will decrease sharply from 81% (2 min maximum waiting time) to 29% (13 min maximum waiting time). However, if the system does not accept ridesharing strategy, the unoccupancy rate is still high (54%) although the passenger can wait for a long time (15 min). The results show that a ridesharing strategy can significantly improve the efficiency of the system.

Table 4.2 Relationship between the maximum waiting time and unoccupancy rate

	Maximum Waiting Time (min)													
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Non-Ridesharing	88%	85%	83%	80%	77%	76%	72%	69%	65%	63%	59%	56%	55%	54%
Ridesharing	81%	72%	68%	59%	50%	45%	40%	36%	32%	31%	31%	29%	29%	29%
Decrease Rate	8%	15%	18%	26%	35%	41%	44%	48%	51%	51%	47%	48%	47%	46%

4.4.2 Fixed Fleet Size

In this section, we create different scenarios with fixed fleet size [300, 400, 500, 600, 700, 800, 900, 1000] for analyzing the influence of ridesharing. In the operation of the ridesharing simulation model, we will generate the service vehicle continuously until the maximum number of vehicles reaches the preset fleet size. As we know that the maximum waiting time can significantly influence the system fleet size, we set the maximum waiting time as 4 minutes for each scenario. When the fleet size is less than the threshold, if there is no available service vehicle for the request, the system will generate a new vehicle at the location of the request origin. When the fleet size reaches the threshold, if there is no available service vehicle for the request, the system will assign the nearest no-load vehicle for the request. After the fleet size is satisfied in the system, the passenger will request the vehicles in the model until all the requests are finished. We record all the trajectories of passengers and vehicles for data analysis.

Figure 4.6 (A) illustrates the average waiting time of passengers with different fleet size. From the figure, it indicates that the ridesharing strategy can significantly reduce the average waiting time of passengers especially when the fleet size is low. When the fleet size is 300, the average waiting time is 825.18s and the average waiting time is 171.26s. The average waiting time drops sharply when the fleet size changes to 400, which means that 300 vehicles cannot satisfy the demand of the system without a ridesharing system. For the fleet size 500, the system with ridesharing can reach a 2min average waiting time, however, without ridesharing, the passengers have to wait 40s longer for the service vehicle on average.

In addition, when the fleet size increased, the difference of average waiting time between the scenarios with and without ridesharing becomes lower. It indicates that the ridesharing strategy will be more beneficial when the system does not have enough service vehicles. If the

system has enough service vehicles, the ridesharing strategy will no longer be sufficiently dominant. Although the ridesharing system can improve the efficiency of response time of service vehicles by decreasing the average waiting time of passengers, the passengers also need to detour more distance.

Figure 4.6 (B) presents the average travel time of passengers with and without the ridesharing strategy. Since the passengers do not have extra trave distance, the average travel time of passengers are very similar (around 767s) for each fleet size scenario. The average travel time of passenger is higher for the ridesharing scenario. And the greater the fleet size, the lower the average waiting time. We can find that for the fleet size 300, the average travel time for passengers is 869s, and it will decrease to 821s for the scenario with fleet size 800. It means, when the fleet size is low, the vehicle needs to detour more travel distances to finish the request.

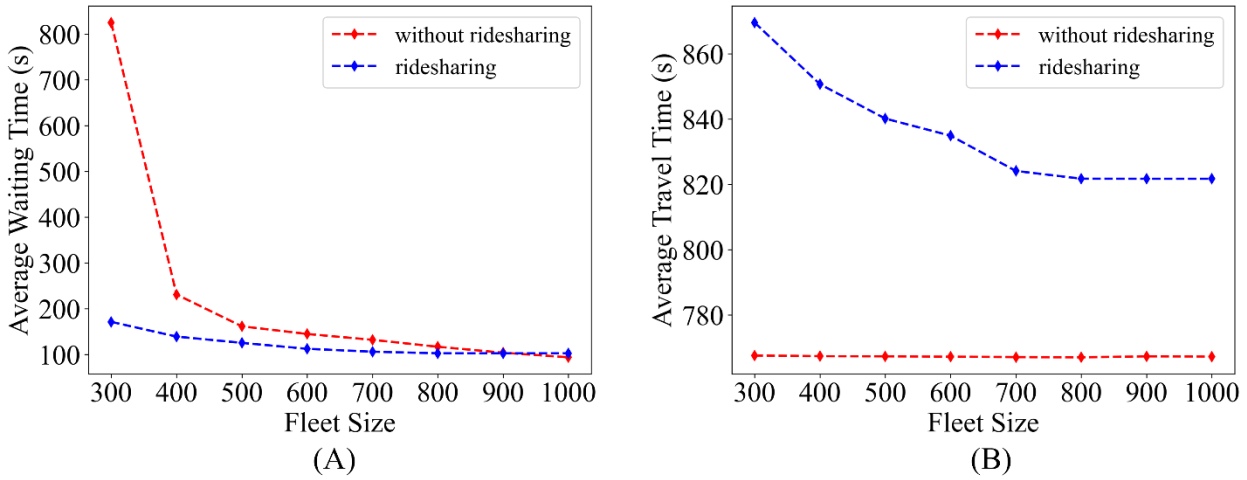


Figure 4.6 Average waiting time (A) and average travel time (B) with different fleet size

A lower fleet size will cause a higher ridesharing rate. The ridesharing rate is defined as the ratio of the passengers choosing ridesharing over the total number of passengers. **Table 4.3** shows the results of ridesharing rate of the passengers with different fleet size. It is notable that, with 300 fleet size in the system, the ridesharing rate can reach to 53.33%, which means more

than half of the passengers will choose the ridesharing vehicles to decrease the waiting time. However, when the system has enough service vehicles, the rate of passengers selected ridesharing will lower. In our system, when the preset maximum waiting time is 4 min, if there are 800 vehicles or more, the ridesharing rate will decrease to 30.53% and more passengers will choose to take the service vehicle alone.

Table 4.3 Ridesharing rate for different fleet size

	Fleet Size							
	300	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
Ridesharing	8717	7284	6628	5752	5120	4996	4996	4996
Non-Ridesharing	7628	9067	9724	10608	11239	11367	11367	11367
Ridesharing Rate	53.33%	44.55%	40.53%	35.16%	31.30%	30.53%	30.53%	30.53%

4.4.3 Vehicle Kilometers Traveled (VKT)

Another benefit of ridesharing is the reduction of vehicle kilometers traveled (VKT). Since different passengers with similar routes can share with one service vehicle, the overlapping distance of these two routes can be saved. Thus, in the operation of the ridesharing simulation system, we recorded the trajectories of each vehicle and calculated the total VKT for both the system with and without ridesharing under different fleet sizes.

Figure 4.7 presents the recorded VKT for the system with and without ridesharing. With different fleet sizes, the VKT shows different patterns with distinguished scenarios. For the system with ridesharing, the VKT shows limited differences with the 8 selected fleet sizes. With the increase of the fleet size, the VKT decreased slowly from 7.45×10^4 km to 7.28×10^4 km. The pattern implies that with the ridesharing strategy, the fleet size only has slight influence on the VKT. However, if the system is without ridesharing, the fleet size can significantly affect the VKT. The results show that with the fleet size of 300, the VKT of the total system will be 1.23×10^5 km, which is 65.6% more than the VKT of ridesharing system. The results reflect that without ridesharing, if the system does not have enough vehicles, the vehicles must travel more

distance to pick up a passenger, which greatly reduces the efficiency of the system. However, the ridesharing system can solve this problem by sharing the vehicles with different passengers, and passengers do not need to wait for the farther service vehicle. When the fleet size is increased, the total VKT will decrease to 7.85 x 10^4 km, which is also 7.8% higher than the VKT of ridesharing system.

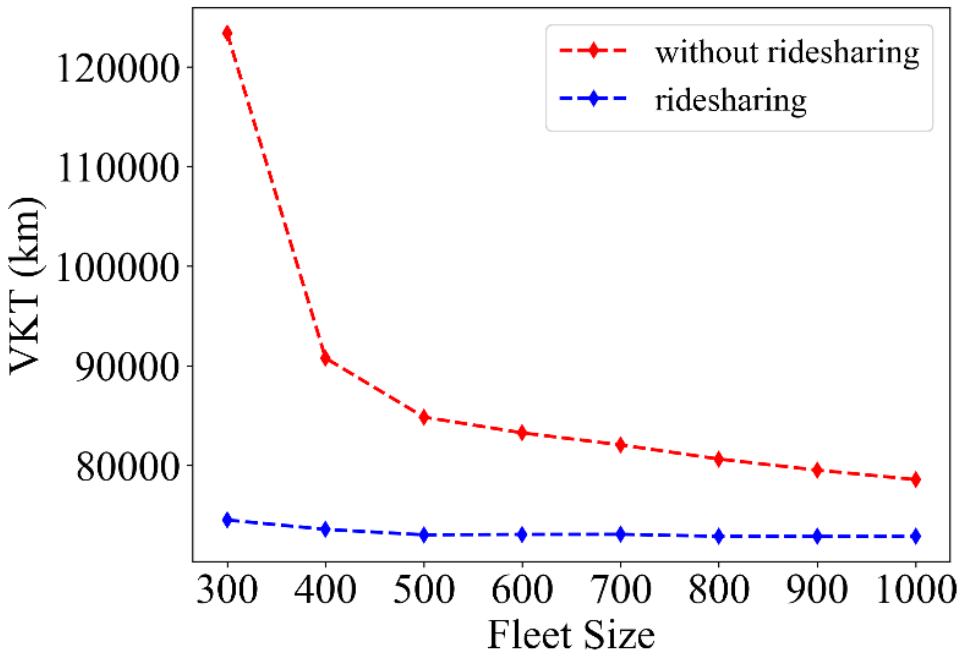


Figure 4.7 VKT with different fleet size

4.4.4 Environment Benefits

The reduction of VKT in the ridesharing system can bring environmental benefits, especially for lowering the emission of greenhouse gas (GHG). Based on [101], transportation is the major source of greenhouse gas emissions. In 2020, around 33% of the total CO_2 emissions and 26 % of the total greenhouse gas emissions in United States are caused by human and goods movement. Ridesharing is one of the solutions to decrease the greenhouse gas emissions caused by transportation. In [101], for each travel mile of passenger vehicles, it will generate 4.03 x 10^-

[4] metric tons CO^2 emissions. And the CO^2 emissions in our system can be calculated as **Table 4.4**.

For different fleet sizes, the total CO_2 emissions in the system are calculated. We compared the CO_2 emissions between the two scenarios under 8 selected fleet sizes. It shows that the CO_2 emissions reduction of the ridesharing system under 300 fleet size compared to the system without ridesharing is 12.3 metric tons per day (which is 39.6% of the total CO_2 emissions of the system without ridesharing). The CO_2 emissions reduction value will change if we consider the different size of the area. For a larger region, the ridesharing strategy will lower CO_2 emissions. The results also reflect that the ridesharing strategy can be more useful when the system's fleet size is lower. It implies that in congested areas or peak hour, the ridesharing strategy can present better performance.

Table 4.4 CO_2 emissions for different fleet sizes

	Fleet Size							
	300	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
Non-Ridesharing	31.08	22.86	21.37	20.97	20.66	20.30	20.03	19.79
Ridesharing	18.77	18.53	18.39	18.40	18.40	18.35	18.35	18.35
CO_2 Emissions Reduction	12.30	4.33	2.98	2.57	2.26	1.95	1.67	1.43

4.5 Conclusions

In recent years, with an increasing number of problems caused by human movements, such as congestion and GHG emissions, ridesharing can be one of the potential solutions to improve the efficiency of the transportation system. In this study, we built a real-time ridesharing system based on an agent-based simulation model to evaluate the impacts of the ridesharing strategy both for its travel and environmental benefits. To make the simulation model closer to the real-world situation, we extract the traffic demand data from Didi ride-hailing service platform and the traffic conditions data using Google Map API. Our model applied a heuristic algorithm to

match the passengers and vehicles in real time under both ridesharing and non-ridesharing scenarios. We tested our model with Didi ride-hailing service data and compared the model performance between the ridesharing and non-ridesharing scenarios.

The major findings of this study are: (1) the system fleet size is related to the preset maximum waiting time of the passengers. If the passengers can wait more time for the service vehicle, the system can lower the fleet size to satisfy the passengers requests. The results show that the ridesharing system can decrease up to 55% of the fleet size and 51% of the unoccupancy rate compared to the system without ridesharing. (2) with 8 fixed fleet sizes, we compared the performance of the average waiting time and average travel time for both ridesharing and non-ridesharing systems. With a lower fleet size, ridesharing can decrease the waiting time for the passengers. The average waiting time can lower 79.2% with fleet size 300 in our system. (3) ridesharing strategy can also reduce the vehicle kilometers traveled in the system. The ridesharing system can decrease a range from 7.6% to 65.6% of the total vehicle kilometers traveled considering different fleet size in the system. In addition, the reduction of vehicle kilometers traveled can also bring environmental benefits. By calculation, we can find with ridesharing, the study area can reduce a range of 7.2% to 40% CO_2 emissions under the system with different fleet size.

There are also some limitations in this paper. First, we selected a subarea in Beijing for the simulation, which means all the influence reported in this paper are based on the specific area. In the future, we will apply this model to a larger area to provide more valuable insights. Second, the ridesharing strategy will bring associated fare change policies since the passengers are more likely to take a ridesharing vehicle with a discount. Third, we assumed that all the passengers will accept the ridesharing strategy which may enlarge the benefits of ridesharing in

real world applications. In the future, we will improve our study by considering the fare structure, which may provide additional valuable information.

CHAPTER 5: CONCLUSIONS

Previous studies on human mobility have mainly used travel surveys, mobile phone, and social media data which may not provide enough details of individual travel choices. Alternatively, emerging data sources can provide more comprehensive transportation-related information. Such data sources include GPS records and ridesharing platforms that can be used to understand and model individual mobility behavior. These data sources can provide information on travel patterns, trip origins and destinations, and mode of transportation used. By analyzing such data, researchers can identify factors that influence travel behavior and develop models to predict future mobility patterns.

In addition, these data sources can be used to evaluate the impact of ridesharing strategies on individual mobility behavior. For example, by comparing travel behavior before and after the introduction of a ridesharing service, researchers can assess the effect of this service on travel patterns, including mode of transportation, trip frequency, and trip length. This information can be used to improve ridesharing strategies and inform transportation policy decisions.

The dissertation aims to analyze human mobility and activity behavior using emerging data sources such as GPS-based trajectory data and on-demand ride-hailing service data. The goal is to develop new methodologies for understanding and modeling individual mobility behavior, which can be useful for transportation planning, intelligent transportation systems, smart cities, and traffic management. The dissertation also aims to evaluate the impact of ridesharing strategies on transportation systems using on-demand ride-hailing service data. Overall, the aim is to gain insights of human mobility and activity behavior and to inform transportation policy to improve the efficiency of transportation networks.

In summary, this study has focused on three objectives, the first objective is to develop an individual activity generative model using long-term GPS-based individual-level trajectory data. The second objective is to develop an individual travel behavior prediction model using large-scale on-demand ride-hailing service data and the third and final objective is to assess the potential of the ride-sharing system using an agent-based simulation model.

5.1 Summary of Major Results

This study provides key insights on understanding and modelling human mobility and activity behavior and evaluating the potential impacts of a ride-sharing strategy using emerging data sources, such as GPS-based trajectory data and on-demand ride-hailing service data. By achieving the objectives, the study aims to gain a deeper understanding of individual mobility behavior and contribute to the development of transportation policy and management strategies. We have summarized the key findings of the study as follows:

- In the second chapter, we first developed an algorithm that can identify the type of activities based on the POI (Point of Interest) category, start time, and duration of the activity. According to the results, the algorithm performs well in identifying activity types. We also developed an input-output hidden Markov model (IOHMM) to generate individual activity sequences given contextual information. Since individual activity patterns are highly heterogeneous, we trained the IOHMM separately for each individual. Trained with massive individual-level GPS-based trajectory data, the results imply that the developed model can accurately generate individual activity sequences, as evidenced by the comparison of predicted values and ground truth values for both the number and duration of activities. It also suggests that if the activities of an individual are known, the model can accurately predict the corresponding activity locations. This means that the model can generate realistic activity

sequences for individuals and can be used to predict the locations where the activities are likely to occur. The model's explainability is a key advantage over previous methods like neural networks and tree-based machine learning model since it enables us to understand the transition probabilities for different activities based on contextual information such as the time of day and day of the week, which can aid in validating and refining the model for improved accuracy through iteration. These results demonstrate the significant potential of the developed model for use in practical transportation planning applications such as transportation simulation, travel demand estimation, and OD matrix prediction, indicating that it can provide accurate predictions to aid in decision-making and optimization.

- In the third chapter, we employed a novel individual-level mobility prediction model - a supervised learning-based multi-layer hidden Markov model - to predict trip decisions and next ride-hailing service trips for each individual using large ride-hailing service data while accounting for travel purpose heterogeneity. To account for the heterogeneous mobility patterns of ride-hailing service users, we divided individuals into four groups - home-based, work-based, commute-based, and random users - and trained the model separately for each group. Results show that the proposed trip decision model outperforms the logistic regression model, achieving an accuracy of around 65%. The emission parameters accurately capture the patterns of daily trip number, and the mobility generation model works well for home-based, work-based, and commute-based users for origin and destination prediction. We also examined the impact of temporal variation on model accuracy and found that accuracy is higher during peak hours and weekdays for the aforementioned groups. To validate the model's performance, we evaluated entropy and predictability for each individual and found that predictability peaks at around 60%. This study is the first to forecast the mobility behavior of on-demand ride-hailing service users,

which could provide valuable insights for policymakers, urban planners, and other stakeholders in the transportation industry.

- In the fourth chapter, we developed an agent-based simulation model to evaluate potential impacts of a real-time ridesharing strategy on transportation efficiency and environmental sustainability. The use of real-world data from a ride-hailing service platform, the traffic status extracted from Google Map API and a heuristic algorithm to match passengers and vehicles in real-time add realism to the simulation model. The results show that if the passengers are willing to wait for a longer period of time for the service vehicle, the ridesharing system can lower the fleet size and unoccupancy rate while still meeting the passengers' requests. Specifically, the results showed that the ridesharing system can decrease up to 55% of the fleet size and 51% of the unoccupancy rate compared to the system without ridesharing. In addition, the ridesharing system can reduce the total vehicle kilometers traveled by a range of 7.6% to 65.6% depending on the fleet size used in the system. Implementing a ridesharing system can also lead to a reduction in CO2 emissions in the study area. Specifically, the reduction in CO2 emissions can range from 7.2% to 40% with different fleet sizes in the system.

5.2 Limitations and Future Research Directions

While this dissertation offers important contributions towards developing a generative model for individual activity sequence, a predictive model for individual mobility behavior, and evaluating the environment and transportation impacts of real-time ridesharing system, there are also some limitations that should be addressed in future research.

- In the individual activity identification process, one limitation is the need for ground truth data to verify the activity identification algorithm used in the model. Obtaining more real-world individual activity data could improve the algorithm's performance in the future.

Another limitation is the model's poor performance in predicting personal activities due to the lack of characteristics of the individuals. Since the heterogeneity of individual travel patterns will increase the randomness of their activities, it will be harder to predict individual activity sequence with high accuracy. Future research could incorporate individual characteristics such as age, gender, and occupation to improve the model's accuracy.

- For individual mobility behaviors predictive model, the dissertation focuses primarily on on-demand ride-hailing services, but the transportation system is complex and involves various modes of transportation. Future research can focus on integrating the findings with other modes of transportation to provide a more comprehensive understanding of the transportation system.

- In the study of evaluating the impacts of a ride-sharing system, first, the study relies on specific data sources, such as a ride-hailing service platform and a mapping API, which may not be representative of all transportation systems. Future research should consider using additional data sources to validate and expand upon the findings of this study. Second, the study focuses on a specific geographic area and may not be generalizable to other regions with different traffic patterns, infrastructure, and travel behaviors. Future research should test the applicability of the proposed methodology in different regions and contexts. Third, the study employs a computationally intensive agent-based simulation model to evaluate the impacts of ridesharing on travel and environmental benefits. Future research should explore alternative methodologies that can achieve similar results with less computational burden. Finally, the study assumes that all passengers will accept the ridesharing strategy. In reality, some passengers may not be willing to share a vehicle with strangers or may have other preferences. Future research

should investigate the user acceptance of ridesharing strategies and incorporate this into the simulation model.

Despite the limitations, this dissertation contributes to the advancement of understanding and modeling individual activity and mobility behavior, as well as evaluating the impacts of a real-time ridesharing systems through the utilization of emerging data sources. Understanding and modeling individual activity and mobility behavior can lead to a better understanding of how people move and interact with their environment. This knowledge can be used to design more effective transportation systems, including public transit, bike share, and ridesharing services. By evaluating the impacts of ridesharing systems, we can identify potential benefits, such as reduced traffic congestion and improved air quality, as well as any negative impacts, such as increased traffic in certain areas or reduced use of public transit. This information can help transportation planners make informed decisions about how to design and implement ridesharing systems to maximize their benefits while minimizing any negative impacts.

REFERENCES

1. Tian, G., J. Wu, and Z. Yang, *Spatial pattern of urban functions in the Beijing metropolitan region.* Habitat International, 2010. **34**(2): p. 249-255.

2. Zhang, J., et al., *Data-driven intelligent transportation systems: A survey.* IEEE Transactions on Intelligent Transportation Systems, 2011. **12**(4): p. 1624-1639.

3. Pan, G., et al., *Trace analysis and mining for smart cities: issues, methods, and applications.* IEEE Communications Magazine, 2013. **51**(6): p. 120-126.

4. Goh, S., et al., *Modification of the gravity model and application to the metropolitan Seoul subway system.* Physical Review E, 2012. **86**(2): p. 026102.

5. Wolf, J., R. Guensler, and W. Bachman, *Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data.* Transportation Research Record, 2001. **1768**(1): p. 125-134.

6. Wu, J., et al., *Automated time activity classification based on global positioning system (GPS) tracking data.* Environmental Health, 2011. **10**(1): p. 1-13.

7. Jiang, B. and Y. Fei, *Traffic and vehicle speed prediction with neural network and hidden markov model in vehicular networks*, in *2015 IEEE intelligent vehicles symposium (IV)(pp. 1082-1087)*. 2015, IEEE. p. 1082-1087.

8. Van Hinsbergen, C., et al., *Bayesian neural networks for the prediction of stochastic travel times in urban networks.* IET intelligent transport systems, 2011. **5**(4): p. 259-265.

9. Polson, N.G. and V.O. Sokolov, *Deep learning for short-term traffic flow prediction.* Transportation Research Part C: Emerging Technologies, 2017. **79**: p. 1-17.

10. Wang, J., et al. *Traffic speed prediction and congestion source exploration: A deep learning method*. in *2016 IEEE 16th international conference on data mining (ICDM)(pp. 499-508)*. 2016. IEEE.

11. Ma, X., et al., *Learning traffic as images: A deep convolutional neural network for large-scale transportation network speed prediction.* Sensors, 2017. **17**(4): p. 818.

12. Huang, S.-H., *An application of neural network on traffic speed prediction under adverse weather conditions*. 2003: The University of Wisconsin-Madison.

13. Rashidi, T.H., et al., *Exploring the capacity of social media data for modelling travel behaviour: Opportunities and challenges.* Transportation Research Part C: Emerging Technologies, 2017. **75**: p. 197-211.

14. Manfredini, F., et al., *Treelet decomposition of mobile phone data for deriving city usage and mobility pattern in the Milan urban region.* Advances in complex data modeling and computational methods in statistics, 2015: p. 133-147.

15. Banister, D., *The sustainable mobility paradigm.* Transport policy, 2008. **15**(2): p. 73-80.

16. Alisoltani, N., L. Leclercq, and M. Zargayouna, *Can dynamic ride-sharing reduce traffic congestion?* Transportation research part B: methodological, 2021. **145**: p. 212-246.

17. Yu, B., et al., *Environmental benefits from ridesharing: A case of Beijing.* Applied energy, 2017. **191**: p. 141-152.

18. Qiao, Y., et al., *A mobility analytical framework for big mobile data in densely populated area.* IEEE transactions on Vehicular Technology, 2016. **66**(2): p. 1443-1455.

19. Pappalardo, L., et al., *Returners and explorers dichotomy in human mobility.* Nature communications, 2015. **6**(1): p. 8166.

20. Schneider, C.M., et al., *Unravelling daily human mobility motifs.* Journal of The Royal Society Interface, 2013. **10**(84): p. 20130246.

21. Song, C., et al., *Limits of predictability in human mobility.* Science, 2010. **327**(5968): p. 1018-1021.

22. Eagle, N. and A.S. Pentland, *Eigenbehaviors: Identifying structure in routine.* Behavioral ecology and sociobiology, 2009. **63**: p. 1057-1066.

23. Gonzalez, M.C., C.A. Hidalgo, and A.-L. Barabasi, *Understanding individual human mobility patterns.* nature, 2008. **453**(7196): p. 779-782.

24. Calabrese, F., G. Di Lorenzo, and C. Ratti. *Human mobility prediction based on individual and collective geographical preferences*. IEEE.

25. Zhao, Z., H.N. Koutsopoulos, and J. Zhao, *Individual mobility prediction using transit smart card data.* Transportation research part C: emerging technologies, 2018. **89**: p. 19-34.

26. Mo, B., et al., *Individual mobility prediction in mass transit systems using smart card data: An interpretable activity-based hidden Markov approach.* IEEE Transactions on Intelligent Transportation Systems, 2021. **23**(8): p. 12014-12026.

27. Colombo, G.B., et al. *You are where you eat: Foursquare checkins as indicators of human mobility and behaviour*. IEEE.

28. Hasan, S., X. Zhan, and S.V. Ukkusuri. *Understanding urban human activity and mobility patterns using large-scale location-based data from online social media*.

29. Smith, J., *US Census Bureau.* 2014.

30. Gidófalvi, G. and F. Dong. *When and where next: Individual mobility prediction*.

31. Lv, Q., et al., *Big data driven hidden Markov model based individual mobility prediction at points of interest.* IEEE Transactions on Vehicular Technology, 2016. **66**(6): p. 5204-5216.

32. Li, F., et al., *A hierarchical temporal attention-based LSTM encoder-decoder model for individual mobility prediction.* Neurocomputing, 2020. **403**: p. 153-166.

33. Trasarti, R., et al., *Myway: Location prediction via mobility profiling.* Information Systems, 2017. **64**: p. 350-367.

34. Yan, M., et al., *Mobility prediction using a weighted Markov model based on mobile user classification.* Sensors, 2021. **21**(5): p. 1740.

35. Etter, V., et al., *Where to go from here? Mobility prediction from instantaneous information.* Pervasive and Mobile Computing, 2013. **9**(6): p. 784-797.

36. Bhattacharya, A. and S.K. Das. *LeZi-update: An information-theoretic approach to track mobile users in PCS networks*. in *Proceedings of the 5th annual ACM/IEEE international conference on Mobile computing and networking*. 1999.

37. Gidófalvi, G. and F. Dong. *When and where next: Individual mobility prediction*. in *Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, pp. 57-64*. 2012.

38. Calabrese, F., G. Di Lorenzo, and C. Ratti. *Human mobility prediction based on individual and collective geographical preferences*. in *13th international IEEE conference on intelligent transportation systems*. 2010. IEEE.

39. Song, H., et al., *Smart cities: foundations, principles, and applications*. 2017: John Wiley & Sons.

40. Expósito-Izquierdo, C., A. Expósito-Márquez, and J. Brito-Santana, *Mobility as a Service.* Smart cities: Foundations, principles, and applications, 2017: p. 409-435.

41. Furuhata, M., et al., *Ridesharing: The state-of-the-art and future directions.* Transportation Research Part B: Methodological, 2013. **57**: p. 28-46.

42. Huang, Z., et al., *Modeling real-time human mobility based on mobile phone and transportation data fusion.* Transportation research part C: emerging technologies, 2018. **96**: p. 251-269.

43. Hasan, S., X. Zhan, and S.V. Ukkusuri. *Understanding urban human activity and mobility patterns using large-scale location-based data from online social media*. in *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*. 2013.

44. Jurdak, R., et al., *Understanding human mobility from Twitter.* PloS one, 2015. **10**(7): p. e0131469.

45. Hasan, S., et al., *Spatiotemporal patterns of urban human mobility*. Journal of Statistical Physics, 2013. **151**: p. 304-318.

46. Chen, D.Q., et al., *Evaluating and Diagnosing Road Intersection Operation Performance Using Floating Car Data.* Sensors, 2019. **19**(10).

47. Peng, C.B., et al., *Collective Human Mobility Pattern from Taxi Trips in Urban Area.* Plos One, 2012. **7**(4).

48. Stopher, P.R. and S.P. Greaves, *Household travel surveys: Where are we going?* Transportation Research Part A: Policy and Practice, 2007. **41**(5): p. 367-381.

49. Axhausen, K.W., et al., *Observing the rhythms of daily life: A six-week travel diary*. 2002. **29**(2): p. 95-124.

50. Chen, C., et al., *The promises of big data and small data for travel behavior (aka human mobility) analysis.* Transportation research part C: emerging technologies, 2016. **68**: p. 285-299.

51. Ma, X., et al., *Understanding commuting patterns using transit smart card data.* Journal of Transport Geography, 2017. **58**: p. 135-145.

52. Sohn, J., *Are commuting patterns a good indicator of urban spatial structure?* Journal of Transport geography, 2005. **13**(4): p. 306-317.

53. Brockmann, D., L. Hufnagel, and T. Geisel, *The scaling laws of human travel.* Nature, 2006. **439**(7075): p. 462-465.

54. Calabrese, F., et al., *Understanding individual mobility patterns from urban sensing data: A mobile phone trace example.* Transportation research part C: emerging technologies, 2013. **26**: p. 301-313.

55. Hawelka, B., et al., *Collective prediction of individual mobility traces for users with short data history.* PloS one, 2017. **12**(1).

56. Chon, Y., et al. *Evaluating mobility models for temporal prediction with high-granularity mobility data*. in *2012 IEEE International Conference on Pervasive Computing and Communications*. 2012. IEEE.

57. Smith, G., et al. *A refined limit on the predictability of human mobility*. in *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. 2014. IEEE.

58. Lu, X., et al., *Approaching the limit of predictability in human mobility.* Scientific reports, 2013. **3**(1): p. 2923.

59. Qin, S.-M., et al., *Patterns, entropy, and predictability of human mobility and life.* PloS one, 2012. **7**(12).

60. Austin, D., et al., *Regularity and predictability of human mobility in personal space.* PloS one, 2014. **9**(2).

61. Qian, W., K.G. Stanley, and N.D. Osgood. *The impact of spatial resolution and representation on human mobility predictability*. in *International Symposium on Web and Wireless Geographical Information Systems*. 2013. Springer.

62. Ikanovic, E.L. and A. Mollgaard, *An alternative approach to the limits of predictability in human mobility.* EPJ Data Science, 2017. **6**(1): p. 12.

63. Zhao, K., et al., *Explaining the power-law distribution of human mobility through transportationmodality decomposition.* Scientific reports, 2015. **5**(1): p. 1-7.

64. Kang, C., et al., *Intra-urban human mobility patterns: An urban morphology perspective.* Physica A: Statistical Mechanics and its Applications, 2012. **391**(4): p. 1702-1717.

65. Liou, S.-C. and H.-C. Lu. *Applied neural network for location prediction and resources reservation scheme in wireless networks*. in *International Conference on Communication Technology Proceedings, 2003. ICCT 2003.* 2003. IEEE.

66. Akoush, S. and A. Sameh. *Mobile user movement prediction using bayesian learning for neural networks*. in *Proceedings of the 2007 international conference on Wireless communications and mobile computing*. 2007.

67. Zong, F., et al., *Trip destination prediction based on multi-day GPS data.* Physica A: Statistical Mechanics and its Applications, 2019. **515**: p. 258-269.

68. Asahara, A., et al. *Pedestrian-movement prediction based on mixed Markov-chain model*. in *Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems*. 2011.

69. Yin, M., et al., *A generative model of urban activities from cellular data.* IEEE Transactions on Intelligent Transportation Systems, 2017. **19**(6): p. 1682-1696.

70. Hasan, S. and S.V. Ukkusuri, *Reconstructing activity location sequences from incomplete check-in data: a semi-markov continuous-time bayesian network model.* IEEE Transactions on Intelligent Transportation Systems, 2017. **19**(3): p. 687-698.

71. Cui, Y., Q. He, and A. Khani, *Travel behavior classification: an approach with social network and deep learning.* Transportation research record, 2018. **2672**(47): p. 68-80.

72. Gong, L., et al., *Deriving personal trip data from GPS data: A literature review on the existing methodologies.* Procedia-Social and Behavioral Sciences, 2014. **138**: p. 557-565.

73. Ke, J., et al., *Short-term forecasting of passenger demand under on-demand ride services: A spatio-temporal deep learning approach.* Transportation research part C: Emerging technologies, 2017. **85**: p. 591-608.

74. Chen, X.M., M. Zahiri, and S. Zhang, *Understanding ridesplitting behavior of on-demand ride services: An ensemble learning approach.* Transportation Research Part C: Emerging Technologies, 2017. **76**: p. 51-70.

75. Fano, R.M., *Transmission of information: A statistical theory of communications.* American Journal of Physics, 1961. **29**(11): p. 793-794.

76. Gidófalvi, G. and F. Dong. *When and where next: Individual mobility prediction*. in *Proceedings of the First ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems*. 2012.

77. Song, L., et al., *Evaluating next-cell predictors with extensive Wi-Fi mobility data.* IEEE transactions on mobile computing, 2006. **5**(12): p. 1633-1649.

78. Yang, Y.X., et al., *Limits of Predictability in Commuting Flows in the Absence of Data for Calibration.* Scientific Reports, 2014. **4**.

79. Center, W.S.D., *http://data.cma.cn/site/index.html*.

80. Goulet-Langlois, G., H.N. Koutsopoulos, and J. Zhao, *Inferring patterns in the multi-week activity sequences of public transport users.* Transportation Research Part C: Emerging Technologies, 2016. **64**: p. 1-16.

81. McLachlan, G.J. and T. Krishnan, *The EM algorithm and extensions*. 2007: John Wiley & Sons.

82. Ford, H.J., *Shared autonomous taxis: Implementing an efficient alternative to automobile dependency.* Princeton University, 2012.

83. Zachariah, J., et al., *Uncongested mobility for all: A proposal for an area wide autonomous taxi system in New Jersey*. 2014.

84. Barth, M. and M. Todd, *Simulation model performance analysis of a multiple station shared vehicle system.* Transportation Research Part C: Emerging Technologies, 1999. **7**(4): p. 237-259.

85. Galland, S., et al., *Multi-agent simulation of individual mobility behavior in carpooling.* Transportation Research Part C: Emerging Technologies, 2014. **45**: p. 83-98.

86. Fagnant, D.J. and K.M. Kockelman, *The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios.* Transportation Research Part C: Emerging Technologies, 2014. **40**: p. 1-13.

87. Lokhandwala, M. and H. Cai, *Dynamic ride sharing using traditional taxis and shared autonomous taxis: A case study of NYC.* Transportation Research Part C: Emerging Technologies, 2018. **97**: p. 45-60.

88. Martinez, L.M. and J.M. Viegas, *Assessing the impacts of deploying a shared self-driving urban mobility system: An agent-based model applied to the city of Lisbon, Portugal.* International Journal of Transportation Science and Technology, 2017. **6**(1): p. 13-27.

89. Ferguson, E., *The rise and fall of the American carpool: 1970–1990.* Transportation, 1997. **24**(4): p. 349-376.

90. Liu, X., et al., *A passenger-to-driver matching model for commuter carpooling: Case study and sensitivity analysis.* Transportation Research Part C: Emerging Technologies, 2020. **117**: p. 102702.

91. Shaheen, S., A. Cohen, and M. Jaffee, *Innovative mobility carsharing outlook. Transportation Sustainability Research Center.* University of California, Berkeley. Retrieved December, 2016. **19**: p. 2016.

92. Qian, X., et al., *Optimal assignment and incentive design in the taxi group ride problem.* Transportation Research Part B: Methodological, 2017. **103**: p. 208-226.

93. Fagnant, D.J. and K.M. Kockelman, *Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas.* Transportation, 2018. **45**(1): p. 143-158.

94. Zhang, W., et al., *Exploring the impact of shared autonomous vehicles on urban parking demand: An agent-based simulation approach.* Sustainable Cities and Society, 2015. **19**: p. 34-45.

95. Fagnant, D.J., K.M. Kockelman, and P. Bansal, *Operations of shared autonomous vehicle fleet for Austin, Texas, market.* Transportation Research Record, 2015. **2563**(1): p. 98-106.

96. Loeb, B., K.M. Kockelman, and J. Liu, *Shared autonomous electric vehicle (SAEV) operations across the Austin, Texas network with charging infrastructure decisions.* Transportation Research Part C: Emerging Technologies, 2018. **89**: p. 222-233.

97. Vosooghi, R., et al., *Shared autonomous vehicle simulation and service design.* Transportation Research Part C: Emerging Technologies, 2019. **107**: p. 15-33.

98. Bonabeau, E., *Agent-based modeling: Methods and techniques for simulating human systems.* Proceedings of the national academy of sciences, 2002. **99**(suppl_3): p. 7280-7287.

99. Reynolds, C.W. *Flocks, herds and schools: A distributed behavioral model*. 1987.

100. Liu, J., et al., *Tracking a system of shared autonomous vehicles across the Austin, Texas network using agent-based simulation.* Transportation, 2017. **44**(6): p. 1261-1278.

101. Ma, J., et al., *Designing optimal autonomous vehicle sharing and reservation systems: A linear programming approach.* Transportation Research Part C: Emerging Technologies, 2017. **84**: p. 124-141.