

# Understanding the Regular Travel Behavior of Private Vehicles: An Empirical Evaluation and A Semi-supervised Model

Chenxi Liu<sup>1</sup>, Jiawei Cai<sup>1</sup>, Dong Wang\* *Member, IEEE*, Jiaxin Tang, Lei Wang, Huiling Chen, Zhu Xiao

**Abstract**—The mining of private vehicle trajectories can be used to establish the regular travel behavior of private vehicles and provide novel solutions to address problems related to urban traffic. With the help of automotive sensors in private vehicles, we can obtain a considerable amount of trajectory data on connected private vehicles. We herein assessed the regular travel behavior from connected private vehicle trajectory data. First, we explored the travel features of private vehicles through an empirical evaluation of move-and-stay frequency, time, duration, and distance. Second, we acquired the fitting distribution of private vehicle move-and-stay features based on the sum of the squared errors. We then proposed a move-and-stay feature based detection (MSFD) method, i.e., a semi-supervised model with incomplete data labels, to identify regular travelers from a massive private vehicle trajectory dataset. Extensive experiments were conducted on the private vehicle trajectory dataset collected from more than 68,069 private vehicles in Shenzhen, China, from June 1, 2016 to August 31, 2016. We compared the proposed MSFD with the existing baselines. The quantitative experimental results demonstrated that the MSFD outperformed these baselines.

**Index Terms**—Private vehicle, regular traveler, trajectory data mining, semi-supervised learning.

## I. INTRODUCTION

The world has recently witnessed an increasing growth in the number of private vehicles. For instance, according to the Annual Statistical Bulletin of China<sup>1</sup>, by the end of 2020, the number of civilian vehicles reached 280 million, 243 million of which were private vehicles (86% of the total number of civilian vehicles). In Europe, the Directorate-General for Energy and Transport of the European Commission estimated that the percentage of private vehicle traffic activities will remain at above 70% until 2030 [17]. In this context, private vehicles constitute an essential part of urban transportation [15]. Although the increasing number of private vehicles has brought convenience to users, it has also led to a number of problems such as traffic jams, limited parking spots, and pollution.

Fortunately, with the recent advances in information and communications technology (ICT), a massive number of connected private vehicle trajectory data that keep track of human travel behavior can be recorded [9]. Mining private vehicle trajectories to establish the regular travel behavior of private

vehicle users can provide a novel perspective to alleviate urban traffic problems [12]. For example, by analyzing regular travel behavior, we can better understand the traffic tide phenomenon, providing empirical support for policymakers to control traffic flows. We can also derive hot regions through the mining of regular travel behavior. This is particularly useful for city planners because appropriate city planning, such as parking lot constructions, can significantly alleviate numerous traffic problems.

However, existing studies on the travel behavior of private vehicles are insufficient. Most are based on location-based social network (LBSN) data [4], [34], [35], taxi trip data [3], [16], [26], mobile phone call data [2], [22], [25], and smart card data [7], [14], [20], among other data types, and there are some deficiencies with using such data. For example, in the case of LBSN data, it is challenging to reflect the stay behavior of users. For instance, mobile phone call data and smart card data are insufficient to characterize the move-and-stay behavior of users [13]. In addition, taxi data can only correspond to the travel behavior of passengers on a current trip, and it is impossible to reflect the long-term travel behavior of fixed users. Private vehicle trajectory data correspond to the long-term travel behavior of individual vehicle owners and can fully reflect their move-and-stay behavior. To date, a few studies have focused on the travel behavior of private vehicles, with most of them focusing on the move behavior of such vehicles [15], [24], [39]. Nevertheless, as shown in Figure 1 (a to f indicate different timestamps), the travel behavior of private vehicles includes not only the move behavior but also the stay behavior. The stay behavior usually represents the travel demand and preferences of the vehicle owners. For

This work was supported in part by the National Natural Science Foundation of China (Grant nos. 61502162, 61702175, and 61772184), the Fund of the State Key Laboratory of Geoinformation Engineering (Grant no. SKLGIE2016-M-4-2), and the Hunan Natural Science Foundation of China (Grant no. 2018JJ2059). (Chenxi Liu and Jiawei Cai contributed equally to this work. Corresponding author: Dong Wang.)

C. Liu, J. Cai, D. Wang, J. Tang, L. Wang, H. Chen, and Z. Xiao are with the College of Computer Science and Electronic Engineering, Hunan University, Changsha, 410082, China. (e-mail: {cxliu, caijw, wangd, txxin, leiwang, chenhl, zhixiao}@hnu.edu.cn).

<sup>1</sup>[http://www.stats.gov.cn/english/PressRelease/202102/t20210228\\_1814177.html](http://www.stats.gov.cn/english/PressRelease/202102/t20210228_1814177.html)

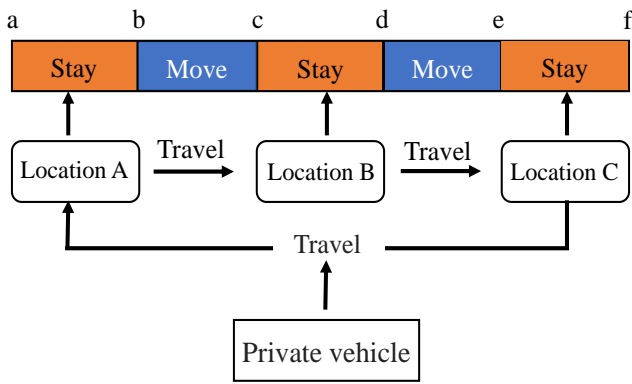


Fig. 1. Travel behavior of private vehicles.

instance, vehicle owners usually stay at their destination for a period after arrival to complete their travel demand and then move to the next location. The stay locations represent the travel demand of the vehicle owners, and the stay duration reflects the travel preference. Therefore, to comprehensively understand the regular travel behavior of private vehicles, it is necessary to conduct an in-depth analysis of both the move and stay behavior.

The limitations of existing studies motivated us to conduct an empirical evaluation of the travel behavior of private vehicles (including move-and-stay behavior). More precisely, we first extract the travel features of private vehicles, including the move-and-stay frequency, time, duration, and distance. Second, we acquire the best fitting distribution and parameters of private vehicle move-and-stay features based on the sum of the squared errors (SSE) and establish that these features follow different probability density distributions between weekdays and weekends. We then propose a move-and-stay feature-based detection (MSFD) model to identify regular travelers from massive private vehicle trajectory data to better understand the regular travel behavior. Specifically, the MSFD is a semi-supervised model with incomplete data labels. The main contributions of this paper are summarized as follows:

- We propose a generic data preprocessing methodology to extract trip data and stay data from connected private vehicle trajectory data.
- We conduct an empirical evaluation of private vehicle travel behavior by analyzing the move-and-stay features extracted from the trip and stay data. The probability density distributions and parameters of the move-and-stay features are acquired through the SSE.
- We propose a semi-supervised model called MSFD to identify regular travelers from massive private vehicle trajectory data under incomplete data scenarios to better understand regular travel behavior.
- We conduct extensive experiments on real-world private vehicle trajectory datasets in Shenzhen, China. The experimental results verify the rationality of the extracted features as the basis of regular traveler detection and the effectiveness of MSFD. We compare the proposed MSFD model with other baselines, and the results of a quantitative experiment demonstrate that the MSFD outperforms

these baselines. In addition, our private vehicle trajectory dataset has been made publicly available<sup>2</sup>.

The remainder of this article is organized as follows. Section II presents related studies on travel behavior. Section III provides basic definitions and introduces some probability distribution functions of private vehicle travel. Following that, the empirical study and analysis results are given in Section IV. Section V introduces the proposed MSFD. Experiments and further analysis are presented in Section VI. Finally, we conclude this article in Section VII with a brief discussion on future areas of research.

## II. RELATED WORK

**LBSN data.** The availability of LBSN data has motivated many researchers to analyze travel behavior. Through an empirical study conducted on a massive check-in dataset collected from Foursquare's Swarm app, the authors in [34] established that travelers' preferences for the location type change with the characteristics of destination city. Hasnat et al. [8] proposed a data mining framework to understand the travel behavior of tourists from social network data. In [35], the authors proposed a visual-analytics framework to model Twitter users' move data and summarized the spatiotemporal travel patterns of users. To mine human travel patterns from geotagged posts, Comito et al. [4] formulated the user move behavior as a sequence of consecutive visits to locations where tweets were posted. LBSN data can record an individual user's location and time. However, the inability to reflect the stay duration, which represents personal preference and travel demand, limits the effectiveness of travel behavior analysis using LBSN data.

**Mobile phone call data.** Some researchers have analyzed human mobility patterns through mobile phone data. In [2], the authors proposed a lifestyle-based clustering to identify users who exhibit similar patterns of mobility behavior; in addition, they developed an entropy-based clustering to measure the homogeneity of mobility patterns within clusters of users. The authors in [22] developed a behavior-oriented time segmentation method for mining individualized time-dependent behavior regularities of mobile phone users. In [25], a classification method for extracting weekly travel patterns from mobile phone call data was proposed. In [38], the authors evaluated the representativeness of mobile phone call data in human mobility characterization from both an individual perspective and a collective perspective. In [29], the authors demonstrated that, if the data are not correctly processed, the statistical regularity of human travel patterns will be overestimated. Mobile phone call data can reflect the long-term travel behavior of users. However, it is difficult to accurately record the travel intention information, such as the destination and move time of a trip. Hence, analyzing travel behavior using mobile phone call data is unsuitable.

**Smart card data.** A considerable number of studies have been conducted to analyze travel behavior using smart card data. The authors in [7] used smart card data to represent the travel behavior of an individual user as a sequence of travel

<sup>2</sup><https://github.com/HunanUniversityZhuXiao/PrivateCarTrajectoryData>

events and proposed a method based on the entropy rate to measure the travel regularity of users. In [20], the authors incorporated the inner-restricted fuzzy C-means clustering and non-negative tensor factorization to predict regional mobility patterns. In [18], the authors extracted mobility patterns by clustering smart card data at the station and passenger levels. In [14], a spectral clustering method was employed to reflect the spatial interactions with other regions and the temporal mobility patterns. In [37], statistical-based and unsupervised cluster-based methods were introduced to analyze the travel patterns of individual users. Smart card data correlate with an individual user's long-term travel behavior. Many researchers use these data to study travel behavior. However, the start and destination of a trip, which reflect the purpose of specific travel behavior, cannot be recorded; thus, it is not the most suitable method for travel behavior analysis.

**Taxi trip data.** An increasing number of researchers are studying travel behavior using taxi trip data. In [3], a novel mobility pattern embedding measure was developed to learn human mobility patterns by modeling sequential, personal, and temporal factors jointly. In [23], a fine-grained semantic pattern extraction system was developed to understand the semantic travel pattern of taxis. In [30], a method that leverages maximum likelihood estimation (MLE) and Bayesian information criterion (BIC) was proposed to fit the patterns of human mobility using taxi trip data. In [36], the authors analyzed urban human mobility patterns using a thematic model based on taxi trip data; they set the road segment identification number of pick-up and drop-off records as semantic information. In [5], the authors leveraged non-negative tensor factorization and Bayesian HyoTrails to characterize human mobility patterns. In [16], the authors developed a clustering algorithm to understand urban mobility using taxi trip data. Taxi trip data contain the pick-up and drop-off locations and the corresponding trips, and as such, they reflect the spatiotemporal features of travel behavior. However, such data cannot correlate with the long-term travel behavior of a specific individual user. Most travel behavior analysis using taxi travel data are based on crowd taxi users, and it is challenging to reflect the travel intention of users.

**Private vehicle trajectory data.** Private vehicle trajectory data correlate with the long-term travel behavior of an individual user. The ability to reflect the move-and-stay information underlying the travel behavior makes it suitable to mine regular travel behavior. Many researchers have recently studied human mobility using private vehicle data, mainly based on two methods. The first method involves data analysis using statistical methods. The authors in [1] analyzed extensive Global Positioning System (GPS) data and statistical regularities for path lengths, activity downtimes, and degrees of activity. In [19], the authors discovered two distinct groups: returners and explorers based on data collected from mobile phones and GPS. Gallotti et al. [6] illustrated the limits of purely descriptive models and provided a mechanistic view of human mobility. The second method involves data mining based on machine learning. In [27], the authors used the kernel density estimation (KDE) method to analyze the stop-and-wait behavior. They proposed a 3D-KDE-based model to describe the

TABLE I  
MAIN NOTATIONS AND DESCRIPTIONS

Notations	Descriptions
$G$	GPS trajectory set
$O$	OBD sequence set
$M$	Move behavior set
$m_f$	Move frequency
$t_s$	Move time
$\delta$	Move duration
$d$	Move distance
$S$	Stay behavior set
$s_f$	Stay frequency
$\tau_s$	Stay time
$\zeta$	Stay duration
$aver\_n_{trip}$	Average number of trips per day
$aver\_d_{trip}$	Average move distance per day
$aver\_t_{trip}$	Average move duration
$ent\_time$	Entropy of the move time
$n_{clusters}$	Number of frequent stay locations
$aver\_t_{stop}$	Average stay duration

dynamic spatiotemporal aggregation effect of private vehicles. In [28], a trajectory aggregation detection (TAD) algorithm was proposed to find areas where the private cars appear frequently in a fixed time interval and then analyzed the travel regularity of each individual private car using trajectory clustering. The authors in [33] proposed a transfer learning method, referred to as ERTB herein, to identify regular travel behavior from private car trajectory data. In addition, in [26], the authors proposed a probabilistic generative model with latent variables to characterize the semantic mobility patterns of private vehicles. The authors in [11] introduced a set of machine learning methods that leverage frequently visited locations to explore the travel patterns of individual users through private vehicle trajectory data. Herein, we extract and analyze the move-and-stay behavior behind private vehicle trajectories using a statistical method. Moreover, based on such behavior, a semi-supervised method is proposed to identify scenarios involving regular travelers with incomplete data.

### III. PRELIMINARIES

We summarize the main notations used in this article in Table I.

#### A. Definitions

**Definition 1 (GPS trajectory).** GPS trajectory set  $G = \{G_1, G_2, \dots, G_n\}$ , where  $G_i = \{g_1, g_2, \dots, g_n\}$  represents the trajectory sequence of a private vehicle with id  $i$ .  $g_n = (id, lon_n, lat_n, t_n)$ , where  $lon_n$  and  $lat_n$  denote the longitude and latitude at timestamp  $t_n$ , respectively.

**Definition 2 (OBD sequence).** The on board diagnostics (OBD) sequence mainly collects vehicle driving status data  $O = \{O_1, O_2, \dots, O_n\}$ . The OBD sequence of the vehicle  $i$  at timestamp  $t_n$  is defined as  $O_i = \{o_1, o_2, \dots, o_n\}$ , where  $o_n = (id, v_n, d_n, t_n, R_n)$ ,  $v_n$  and  $d_n$  are the instantaneous speed and direction, respectively. In addition,  $R_n$  is a remark, e.g., OBD start or OBD flameout.

**Definition 3 (Move behavior).** The move behavior of the private vehicle  $i$  is defined using  $M_i =$

TABLE II  
AN EXAMPLE OF GPS AND OBD FUSION DATA

ObjectID	Lon	Lat	Speed	Direct	Remark	GPSTime
42234	113.411935	22.951953	13.52	330	OBD start	2016/8/31 17:29:25

$(id, t_s, slon, slat, t_e, elon, elat, \delta, d)$ . Here,  $slon$  and  $slat$  denote the longitude and latitude of the starting location, respectively, and  $t_s$  is the corresponding timestamp. Moreover,  $elon$  and  $elat$  denote the longitude and latitude of the ending location, respectively,  $t_e$  is the end time,  $\delta$  is the move duration.  $d$  is the move distance, which is measured based on Haversine distance formula, and is defined as (1)–(3):

$$d = 2 \cdot r \cdot \arctan\left(\sqrt{\frac{a}{1-a}}\right), \quad (1)$$

$$a = \sin^2\left(\frac{elat_i - elat_j}{2}\right) + \cos(elat_i) \cos(elat_j) \sin^2\left(\frac{elon_i - elon_j}{2}\right), \quad (2)$$

where  $r$  represents the average radius of the Earth, which is 6371.0088 kilometers.

**Definition 4 (Stay behavior).** The stay behavior of vehicle  $i$  is defined by  $S_i = (id, lon, lat, \tau_s, \tau_e, \zeta)$ . In addition,  $lon$  and  $lat$  denote the longitude and latitude of the stay location, respectively;  $\tau_s$  is the stop timestamp when the vehicle starts to stay;  $\tau_e$  is the restart timestamp when the driver restarts the vehicle; and  $\zeta$  denotes the stay duration.

**Definition 5 (Move frequency).** Move frequency refers to the average number of move behavior per vehicle per day, which is defined as (4):

$$m_f = \sum_{i=1}^D tc_i / D, \quad (3)$$

where  $tc_i$  indicates the number of move behavior on the  $i$ -th day,  $D$  represents the total number of days the vehicle moves.

**Definition 6 (Stay frequency).** Stay frequency refers to the average number of stay behavior per vehicle per day, which is defined as (5):

$$s_f = \sum_{i=1}^D sc_i / D, \quad (4)$$

where  $sc_i$  is the number of stay behavior on the  $i$ -th day.

### B. Private Vehicle Trajectory Data Collection

We collected private vehicle trajectory data through GPS and OBD devices jointly developed by the industry [32]. As depicted in figure 2, a GPS receiver is used to obtain the vehicle's spatiotemporal information, such as longitude, latitude, and timestamp. The interface reader of the OBD module reads the driving information, such as speed, direction, and remark. The subscriber identity module in the GPS sends the collected trajectory information to the data center. The desensitized

trajectory data collected by the GPS and OBD devices is transmitted to the Mapgoo smart cloud, a service platform that helps developers quickly access various terminal devices, such as GPS devices, OBD devices, and cloud recorders. The data collection process was continuous [31]. In this study, we selected 68,069 private vehicles in Shenzhen, China, from June 1, 2016, to August 31, 2016.

## IV. EMPIRICAL EVALUATION

As shown in Figure 3, we first preprocessed the data. We then conducted an empirical evaluation based on the processed private car trajectory data.

### A. Data Preprocessing

A trajectory dataset usually contains noise, which refers to position data that deviates from the actual value owing to the signal of the vehicle-mounted positioning device. We applied heuristic outlier elimination to process the noise in two steps: *i)* we used an outlier detection model to detect the noise in the original data, and *ii)* we set a threshold speed (150 km/h) and calculated the moving average speed of the two locations before and after the current location, and its corresponding time interval within the trajectory. When the average speed was greater than the threshold speed, the location was marked as noise. After traversing the entire trajectory data, all locations marked as noise were eliminated.

We extracted the trip and stay data of a private vehicle from GPS and OBD data [10]. The trip data mainly recorded the

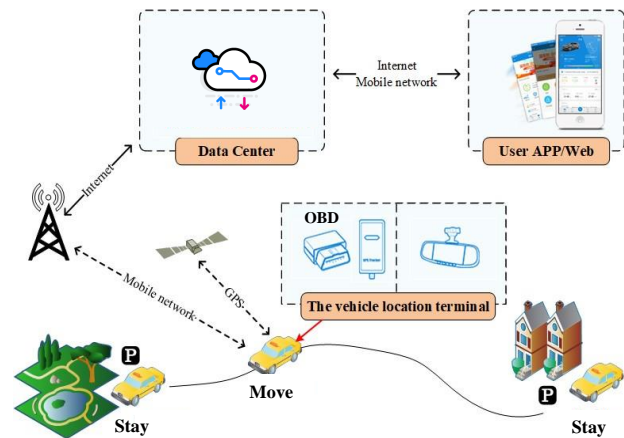


Fig. 2. Process for collecting private vehicle trajectory data.

TABLE III  
AN EXAMPLE OF THE TRIP DATA

ObjectID	StartTime	StartLon	StartLat	StopTime	StopLon	StopLat	Duration	Distance
42234	2016/6/30 22:27:27	114.276773	23.173776	2016/6/30 23:37:36	114.142071	23.242029	1h10min9s	35885



TABLE IV  
AN EXAMPLE OF STAY DATA

ObjectID	StartTime	StopTime	StopLon	StopLat	Stay duration
42234	2016/08/31 08:50:45	2016/08/31 17:29:45	113.41193	22.951953	8h39min

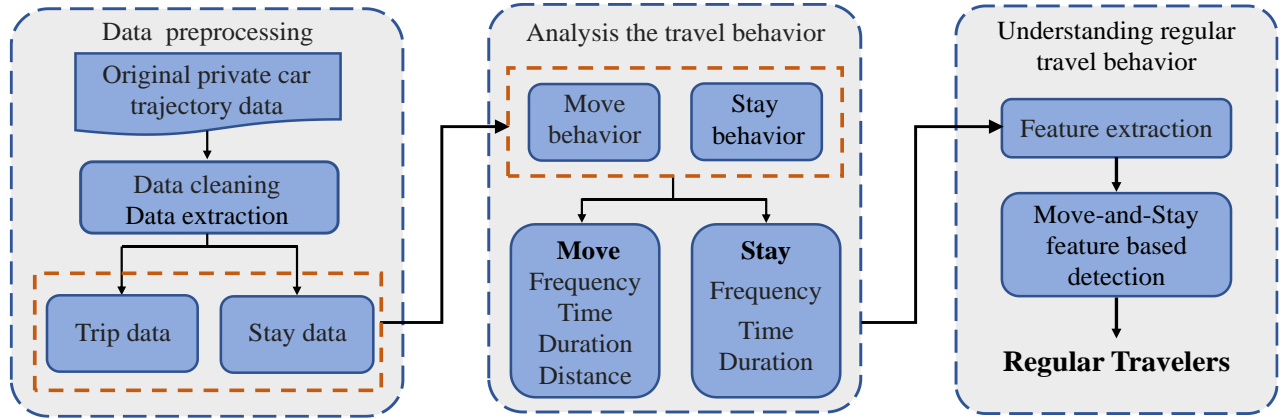


Fig. 3. Proposed framework.

moving behavior of the vehicles, whereas the stay data mainly described the stay behavior of the vehicles. The data extraction process is as follows:

(1) Data filtering: we set a research area  $R$  and period  $T$ . We filtered out the trajectory data from GPS and extracted the driving state data from OBD in the  $R$  and the period range  $T$ . The area of vehicle activity was set to Shenzhen, where the traffic problem is severe. The period range was set to 3 months of vehicle trajectory data from June 1, 2016, to August 31, 2016.

(2) GPS and OBD data fusion: we obtained  $t$  by looking at when the vehicle engine was started and turned off using the OBD data. We then pinpointed the location recorded at the corresponding time in the GPS data and marked the location as the start location or the flameout location. An example of fused data is shown in Table II.

(3) Trip data extraction: we set the time threshold to  $\delta$  and calculated the time interval between the current and next trips. If the calculated time interval was less than  $\delta$ , the two trips were merged. We then set the end location of the following move location as the starting location of the current trip until the complete trip data were traversed to obtain the full trip data of the vehicle. An example of the extracted trip data is presented in Table III.

(4) Obtaining stay data: according to the trip data that were obtained, the end position ( $elon, elat$ ) of the current trip was designated as the stay location ( $lon, lat$ ). The end time of the current trip  $t_e$  was considered the start time  $\tau_s$  of the stay data. The start time of the next trip  $t+1_s$  was taken as the end time of the stay data,  $\tau+1_e$ . An example of extracted stay data is presented in Table IV.

### B. Probability Distribution and Parameter Selection

Herein, the probability density can be moved or scaled using two parameters (loc and scale), and the original probability density is given by  $p(x) = \frac{p(y)}{scale}$ ,  $y = \frac{x-loc}{scale}$ . Therefore,

in addition to the function parameters, the loc and scale parameters were also considered. SSE is the sum of squares of the errors between the fitted data and the original data of the corresponding points, which is applied to evaluate the quality of the fit:

$$SSE = \sum (Y_i - \hat{Y}_i)^2, \quad (5)$$

where  $Y_i$  is the real data and  $\hat{Y}_i$  is the fitted data. The standard form of the probability density distribution function involved in several critical primary attributes of user travel behavior is presented in Table V.

### C. move Behavior Analysis

1) *Move frequency*: Analyzing the move frequency of private vehicle travel provides a reference for governments to promote low carbon travel and environment monitoring. As shown in Figure 4(a), the horizontal axis represents the average number of trips per vehicle per day, and the vertical axis represents the percentage of vehicles corresponding to the total number of vehicles. On weekdays, most vehicles travel twice a day on average and the owners are likely to be commuters. The average number of vehicles that travel twice a day is the highest on weekends. However, in this case, the trips are most likely to be between home and scenic spots. On both weekdays and weekends, it was observed that the vehicle owners travel at least once, whereas on weekends they are more likely to travel more than four times compared to that on weekdays.

2) *Move time*: The move time is one of the critical parameters for analyzing the travel behavior of private vehicles. Understanding the regularities in the move time, e.g., when private vehicles often leave, and inferring the peak and low travel times of most groups within the city, is extremely important for traffic control. As shown in Figure 4(b), both weekdays and weekends show the same regularities. For instance, between 12 am and 5 am, and after 6 pm, the vehicle travel volume gradually decreased because some owners had

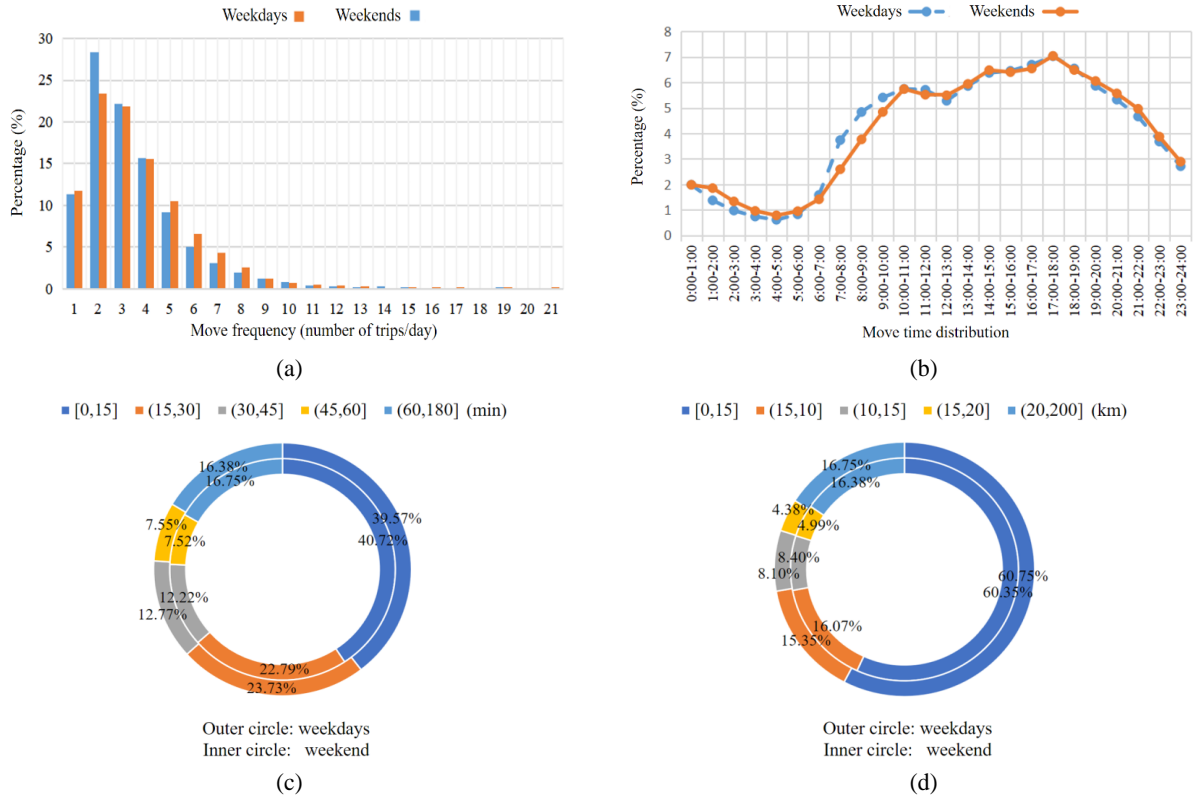


Fig. 4. Move behavior statistics. (a) Move frequency. (b) Move time. (c) Move duration. (d) Move distance.

already returned home to rest during these times. After 5 am, vehicle travel volume steadily increased, and between 10 am and 11 am, the travel volume reached the first peak period.

3) *Move duration*: The traffic conditions of a road network can be inferred by analyzing the distribution of the move duration. In addition, we can recommend a suitable route for vehicle owners based on the historical move duration of private vehicles. As shown in Figure 4(c), most vehicles travel for less than half an hour, accounting for more than 63% of the recorded cases. Simultaneously, combined with the move time analysis, three periods of 10:00–11:00, 17:00–18:00, and 20:00–21:00, which are the peak travel periods, and the two

periods of 4:00–5:00 and 12:00–13:00, which are low periods, were selected. Moreover, the probability density distribution of the move time follows the inverse gamma distribution on weekdays and the exponential Weibull distribution on weekends. The related parameters are presented in Table VI.

4) *Move distance*: The move distance reflects the spatial information related to the move behavior of private vehicles. We understand a vehicle's spatial travel features by analyzing the distribution of the vehicle's move distance. As shown in Figure 4(d), the travel regularities are similar on both weekdays and weekends, and the move distances of private vehicles within 5 km account for more than 60%, indicating

TABLE V  
PROBABILITY DENSITY DISTRIBUTION

Distribution	Expression
Exponential	$f(x) = \lambda e^{-\lambda x}$
Exponential modified Gaussian	$f(x, \lambda, \mu, \sigma) = \frac{\lambda}{2} \exp \left( \frac{\lambda}{2} (2\mu + \lambda\sigma^2 - 2x) \right) \times \operatorname{erfc} \left( -\frac{\mu + \lambda\sigma^2 - 2x}{\sqrt{2}\sigma} \right)$
Invgamma	$f(x, a) = \frac{x^{-a-1}}{\Gamma(a)} \exp \left( -\frac{1}{x} \right), \Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$
Exponweib	$f(x, a, c) = ac [1 - \exp(-x^c)]^{a-1} \times \exp(-x^c) x^{c-1}$
Invgauss Gaussian	$f(x, \mu) = \frac{1}{\sqrt{2\pi}x^3} \exp \left( -\frac{(x-\mu)^2}{2x\mu^2} \right)$
Generalized generalized Pareto	$f(x, c) = (1 + cx)^{-c-1/c}$

TABLE VI  
MOVE DURATION DISTRIBUTION PARAMETERS

Type	Distribution	a	c	Loc	scale	SSE
Weekday	Invgamma	1.3704	/	-1.2479	21.1969	$2.5410^{-5}$
Weekend	Exponweib	2.1066	0.5261	3.01362	10.7948	$1.5910^{-6}$

TABLE VII  
MOVE DISTANCE DISTRIBUTION PARAMETERS

Type	Distribution	$\lambda$	$\mu$	$\sigma$	SSE
Weekday	Exponential	0.1106	/	/	$2.3010^{-4}$
Weekend	Exponential modified Gaussian	0.1054	-0.003	0.0079	$2.6310^{-4}$

TABLE VIII  
STAY DURATION DISTRIBUTION PARAMETERS

Type	Distribution	$\mu$	c	loc	scale	SSE
Weekday	Invgauss Gaussian	6.6810	/	1.3737	23.9266	$1.7010^{-5}$
Weekend	Generalized generalized Pareto	/	2.0310	5.0167	32.2562	$5.7110^{-6}$

that most owners use private vehicles for short trips. The number of vehicles with a move distance of 15–20 km is the lowest, indicating that owners generally do not choose private vehicles as their mode of travel for moderate move distances. Moreover, the probability density distribution of the move distance follows an exponential distribution on weekdays and an exponential modified Gaussian distribution on weekends. The related parameters are shown in Table VII.

#### D. Stay Behavior Analysis

1) *Stay frequency*: Private vehicles have clearer stay behavior than floating vehicles, which underline the staying preferences of the former. Analyzing the stay frequency of private vehicles can help companies accommodate different stay preferences of the owners and reasonably customize their portfolios. As shown in Figure 5(a), the horizontal axis represents the average number of stay positions per vehicle per day. The vertical axis represents the percentage corresponding to the total number of vehicles. On weekdays, users stay twice a day on average, and the two active locations are likely to be work and residence locations. On weekends, users stay once a day on average. However, users are more likely to have two or more stay activities on weekends than on weekdays.

2) *Stay time*: Figure 5(b) shows that weekdays and weekends follow the same regularities. The number of vehicles with stay behavior of between 5 am and 6 am is the lowest, whereas the number of stay behavior between 6 pm and 7 pm is the highest. The stay behavior are generally lower at 9 am than at 12 am, whereas the number of stay behavior occurring from 9 am to 12 pm is lower than those occurring from 2 pm to

8 pm. As a result, we infer that a series of private activities are mainly concentrated at between 9 am and 8 pm with the possibility of activities occurring after 2 pm being relatively high.

3) *Stay duration*: The stay duration provides crucial information that represents urban attractiveness. For example, a more considerable stay time denotes a more substantial attractiveness for a region with the same spatial aggregation. A comprehensive analysis of the duration distribution features provides a reference for understanding the regularities of daily activities such as work, rest, shopping, and leisure. The result of analyzing the overall distribution of the stay behavior durations is shown in Figure 5(c). Regardless of weekdays or weekends, most stay behavior are within 2 h. Only a small portion lasts for more than 2 h, but no more than 8 h. A stay behavior of more than 8 h is rare. Thus, we can infer that most of the daily activities of private vehicles are dining, shopping, and other short-term activities. By contrast, there are fewer long-term activities for over 8 h. At the same time, further analysis of the stay duration reveals that the probability density distribution of the stay duration follows an inverse Gaussian distribution on weekdays and a generalized Pareto distribution on the weekends. The parameters are presented in Table VIII.

#### E. Results

This section discusses the results of the analysis performed in this study. By analyzing the move frequency features of private vehicle users to understand travel needs, we found that private vehicles are most likely to travel twice a day on average regardless of the day of the week. On weekdays, we inferred

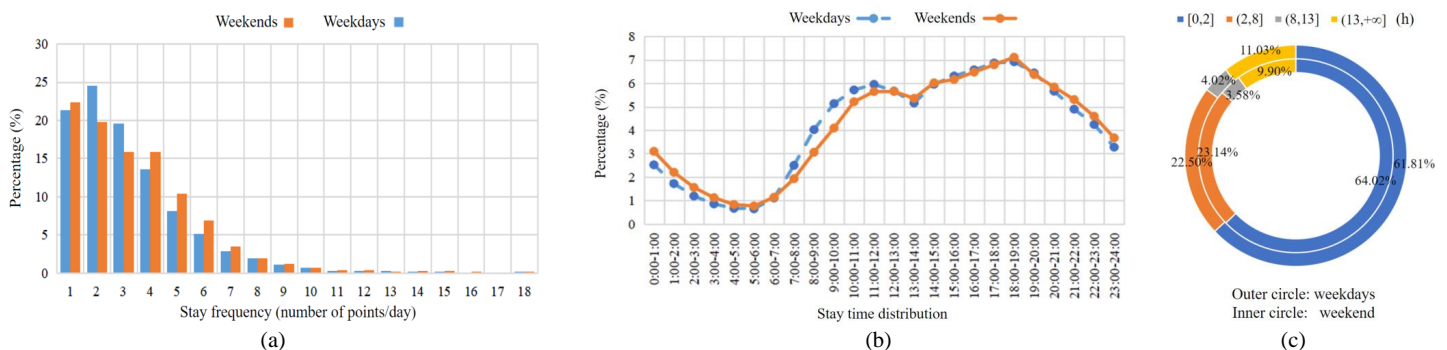


Fig. 5. Stay behavior statistics. (a) Stay frequency. (b) Stay time. (c) Stay duration.

that these two trips could correspond to the commuting needs of the users. On weekends, these two trips are more likely to be return trips between home and entertainment locations. The times when the move first began were collected at 10 am and at 5 pm based on an analysis of the move time. The number of users who started to move at a particular moment represents the number of users traveling at that moment; in addition, we can infer that the two peak travel periods for private vehicle users in Shenzhen are 10 am and 5 pm. During these two periods, the possibility of road traffic congestion is extremely high. Simultaneously, by analyzing the move time and distance to understand the spatiotemporal travel features of the vehicles, we found that private vehicle users usually travel short distances.

Further, each stay behavior corresponds to the vehicle's travel activity. From an analysis of the stay frequency, we found that most users need to perform two activities on a weekday, which are most likely arriving at work and returning home to rest. On weekends, the users tend to have only one activity, which is a noticeable difference from that in weekdays. A possible reason for this is the fact that most users choose to stay at home on weekends. Regarding the stay time, we found that private vehicle users are primarily engaged in a series of private activities between 9 am and 8 pm. Further, the possibility of having activities after 2 pm is relatively high. Work activities are more likely to occur during weekdays, whereas leisure and entertainment activities are more likely to occur on weekends. Furthermore, the stay duration reflects the idea that short-term activities often occur between 2 and 6 pm. On weekends, activities lasting 2-8 h usually occur between 12 am and 6 am. The long-term 8-13 h activities mainly occur between 6 and 8 pm. Long-term activities of more than 13 h generally occur at 6 and 8 pm. We infer that short-term activities such as leisure or dining mainly occur between 2 and 6 pm. More extended activities such as work mainly occur between 12 and 9 am, and long-term rest mainly occurs after 6 pm.

The aforementioned results can provide a reference for formulating urban traffic control strategies, such as obtaining the peak travel period of private urban vehicle users at the beginning of the move and setting traffic diversion measures during peak hours. Because the stay behavior corresponds to the activities, a recommendation can be made according to the travel regularities of a crowd. For example, on weekends, short-term activities mainly occur at 2 pm. Thus, at this time, the decision maker can make advertisement recommendations for entertainment and shopping locations for users. In addition, the extracted probability distributions can be referenced for private vehicle travel behavior.

## V. SEMI-SUPERVISED DETECTION MODEL

We propose a semi-supervised detection model called move-and-stay feature-based detection (MSFD) to identify regular travelers in incomplete data scenarios, as shown in Algorithm 1. Regular travelers refer to vehicle owners who have repeated travel behavior within a specified period and meet the following conditions:

- Vehicle users have a similar start time sequence and end time sequence on each weekday;
- Vehicle trips at similar timestamps have the same start and end locations.

The orange segment in Figure 6(a) represents a vehicle in a moving state, and the green segment represents a vehicle in a stopped state. It can be seen from Figure 6(b) that from June 1 to 4, 2016, the vehicle owners drove from location A to location B at approximately 9 am every day, and then from location B to location A at approximately 10:00 pm. According to their observed travel behavior, they travel between the same locations A and B at similar times. Similarly, private vehicles that do not meet the two aforementioned conditions at these times when traveling belong to an irregular travel group. Therefore, they are called irregular travelers.

We extracted the move-and-stay behavior features from the travel behavior of private vehicles:

1) *Move behavior feature*: The average move frequency per day  $aver\_n_{trip}$ : regular travelers on workdays tend to travel to and from home and work and do other repetitive activities. Thus, the number of round trips constitutes the majority of all trips, and the average number of daily trips of regular travelers is more likely to be greater than one. For irregular travelers, their trips are random because, as compared with regular travelers, the probability that the average number of trips per day will be odd is higher. The two types of private vehicles have different average daily trips, which can be calculated as follows:

$$aver\_n_{trip} = \sum_{i=1}^d n_{trip}(i)/d, \quad (6)$$

where  $n_{trip}(i)$  represents the number of vehicle trips on the  $i$ -th day, and  $d$  represents the total number of days of the trip.

Average move distance per day  $aver\_d_{trip}$ : the travel behavior of regular travelers repeated periodically. Considering the

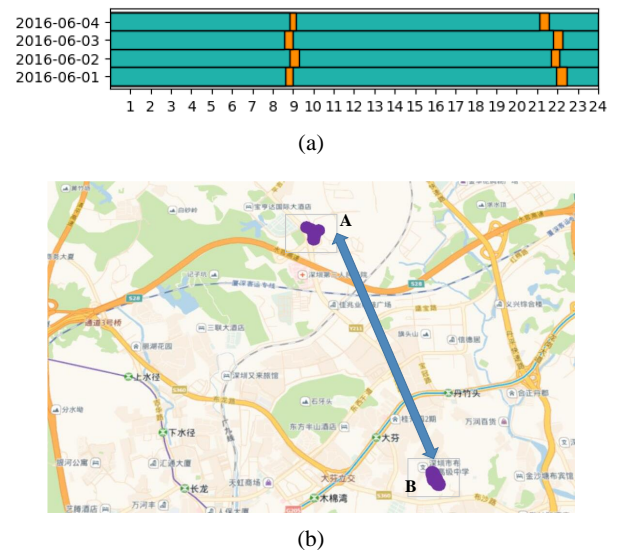


Fig. 6. Spatio-temporal features of regular travelers. (a) Spatial feature. (b) Temporal feature.



cost of traveling, travelers tend to prefer short-distance trips. It was observed that, with regard to travel needs, irregular travelers prioritize travel convenience. Therefore, when these travelers have long-distance travel needs, they choose private vehicles to travel in. As a result, under the same period and travel demands, the average travel distance of irregular travelers has a greater probability of being long as compared to that in the case of regular travelers. This can be calculated as follows:

$$aver\_d_{trip} = \sum_{i=1}^n dist_{trip}(i)/n. \quad (7)$$

Average move duration  $aver\_t_{trip}$ : irregular travelers are more inclined to take long-distance trips than regular travelers. Under the same period and travel demand, the trip duration is longer. The average travel time of the vehicle is calculated by dividing the total travel time by the number of trips:

$$aver\_t_{trip} = \sum_{i=1}^n t_{trip}(i)/n. \quad (8)$$

The entropy of the move time  $ent\_time(u)$ : we use the time information entropy to highlight the degree of mixing of a single start time of a private vehicle. The smaller the information entropy is, the stronger the regularity of the travel start time of the vehicle. We divide the 24 hours of a weekday into 24 time blocks, with an interval of 1 h between each block. In the trip data, each trip is divided into the number of time blocks according to the start time. The proportion of the number of trips in the  $i$ -th time block is  $p_i$ , and the start time information entropy of the vehicle trip is defined as follows:

$$ent\_time(u) = - \sum_{i=1}^{24} p_i \log_2 p_i. \quad (9)$$

2) *Stay behavior feature*: Frequent stay locations  $n_{clusters}$ : we clustered the staying locations of private vehicles. The number of clusters obtained represents the number of frequent staying locations of vehicle owners in a fixed period. To identify the different stay positions contained in the stay location trajectory data, the  $eps$  parameter was set to 0.4 km, and the MinPts parameter was used to control the least MinPts stay locations to form a cluster. To capture the number of locations where vehicles frequently stay, this parameter was set to half the number of days in the cycle. The stay locations corresponding to the formed clusters are frequent areas where vehicle owners visit at least once in 2 days.

Average stay duration  $aver\_t_{stop}$ : the activities conducted by regular travelers who exhibit staying behavior are generally long-term activities such as work and home rest, whereas the activity performed by irregular travelers who exhibit staying behavior is shopping. Such short-term and occasional activities are not routine daily activities. Therefore, the average travel duration features of the two types of private vehicle groups are different and were calculated as follows:

$$aver\_t_{stop} = \sum_{i=1}^n t_{stop}(i)/n, \quad (10)$$

where  $t_{stop}(i)$  represents the stay durations at the  $i$ -th stay location, and  $n$  represents the total number of vehicle stay locations.

To identify regular travelers, we used a small number of labeled private vehicle user samples as supervised information; in other words, we added seed constraints and sampled the partition constraints to divide private vehicle users into groups with different travel intentions. The samples belonging to regular travelers were called positive samples, and the samples belonging to irregular travelers were called negative samples. Then, we assumed that in the labeled data  $X_l$  in the training set, where the positive sample set was  $X_l^1$ , and the negative sample set was  $X_l^0$ . The seed constraints and sample partition constraints added for the training process using labeled samples were as follows:

**Initial centroid constraint.** The initial centroid is the two samples with the farthest distance between the positive and negative samples in the labeled sample set, and the label of the initial centroid does not change; in other words, the group to which the initial centroid belongs does not change. The travel behavior features extracted from  $N$  private vehicle users were given as input to the MSFD model. Because there were six types of extracted move-and-stay behavior features for each user, the Euclidean distance between the two feature vectors  $x_i$  and  $x_j$  was estimated as follows:

$$d(x_i, x_j) = \sqrt{\sum_{n=1}^6 (x_i^n - x_j^n)^2}. \quad (11)$$

The initial centroids  $x_0$  and  $x_0'$  satisfy the following:

$$\begin{cases} x_0 \in X_l^1, x_0' \in X_l^0 \\ d(x_0, x_0') = \max(\forall x_i \in X_l^1, x_j \in X_l^0 d(x_i, x_j)) \end{cases} \quad (12)$$

**Sample division constraint.** Because of the insufficient amount of labeled data, the distance between the sample and the centroid could not be used as the only basis for determining the sample division. We therefore set it in a labeled sample set. Notably, samples with the same label must belong to the same group. Such samples have a *link* correlation. Further, samples with different labels must not belong to the same group; that is, they must have a *not link* correlation. When a sample belongs to unlabeled data, the distance from the sample to the centroid is used as the classification basis. The current sample is labeled, and the type and quantity of the label data in the target cluster must be considered. Therefore, this study defined the sample division to take the group with the smallest  $D(x_i, C_j)$  as the type of sample  $x_i$ :

$$D(x_i, C_j) = \begin{cases} d(x_i, x_0^j) & x_i \notin X_l \\ d(x_i, x_0^j) * \frac{n\_notlink(x_i) - n\_link(x_i)}{N(X_l)} & x_i \in X_l \end{cases}, \quad (13)$$

where  $C_j$  is the target group of the current sample,  $x_0^j$  is the current centroid of the target group,  $n\_notlink(x_i, C_j)$  is the number of samples in the target group that have a *not link* correlation with the current sample,  $n\_link(x_i, C_j)$  is the number of samples in the target group that have a *link*

correlation with the current sample, and  $N(X_l)$  is the number of sample points in the labeled dataset.

The process of identifying regular travelers of the MSFD model is shown in Algorithm 1:

(1) We selected the initial centroid in the dataset according to Equations (11) and (12).

(2) For the unlabeled samples, we calculated the distance to each centroid and classified the samples into the group of centroids with the smallest distance. For the labeled samples in the dataset, we calculated the distance between the sample and each centroid. The difference between the number of samples of *not link* and *link* correlations in each group was taken as the percentage of the total number of labeled samples. The samples were classified into the group with the smallest product of the centroids, as shown in Equation (13).

(3) According to Equation (14), we updated the centroid of the group to be the mean of all the samples.

$$x_0 = \frac{x_1 + x_2 + \dots + x_n}{n}, \quad (14)$$

(4) We repeated steps (2) and (3) until the centroid of each group no longer changes. It was then considered that a stable state had been reached, and the classification result was given as the output.

---

**Algorithm 1** Move-and-stay Feature-based Detection

---

**Input:** Private vehicle sets: labeled sample set  $X_l$  and unlabeled sample set  $X_u$

**Output:** Regular traveler set  $C_0$  and irregular traveler set  $C_1$

```

1: Initial  $C_0 = C_1 = \emptyset$ 
2:  $C = X_l \cup X_u, p = 0, q = 0, dist = 0$ 
3: Randomly select 100 positive samples and 100 negative
   samples from  $X_l$  to form  $X_l^0$  and  $X_l^1$ , and set the initial
   centroid  $x_0 = X_l^0[0], x'_0 = X_l^1[0]$ 
4: for each  $x_i \in X_l^0$  do
5:   for each  $x_j \in X_l^1$  do
6:     Calculate  $d(x_i, x_j)$  according to Equations (11)
       and (12)
7:   if  $d(x_i, x_j) > dist$  then
8:      $p \leftarrow i, q \leftarrow j$ 
9:      $dist \leftarrow d(x_i, x_j)$ 
10:  end if
11: end for
12: end for
13: Obtain the initial centroid as  $x_0 = X_l^0[p], x'_0 = X_l^1[q]$ ,
    append  $x_0$  to  $C_0$ , and append  $x'_0$  to  $C_1$ 
14: for each  $x_i \in C$  do
15:   Calculate  $D(x_i, C_0)$  and  $D(x_i, C_1)$  according to
       Equation (13)
16:   if  $D(x_i, C_0) < D(x_i, C_1)$  then
17:      $x_i$  appends to  $C_0$ 
18:   else
19:      $x_i$  appends to  $C_1$ 
20:   end if
21:   Update  $x_0$  and  $x'_0$  according to Equation (14)
22: end for
```

---

## VI. EXPERIMENTS

### A. Dataset

Herein, we selected 68,069 private vehicles in Shenzhen, China, from June 1, 2016, to August 31, 2016. We randomly extracted the travel behavior data of 1,000 private vehicles on weekdays, 400 of which were marked as private vehicles, including 200 regular travelers. The label was set to  $Y = 1$ . The method of obtaining the label was based on our previous study [11]. For 200 irregular travelers, the label was set to  $Y = 0$ . In training the model, we randomly extracted 100 regular and 100 irregular travelers from the labeled data, and all the other samples were treated as unlabeled data. We extracted the travel behavior features of each vehicle as a sample set. The travel behavior features of each vehicle were used to constitute a feature vector as a sample point in the sample set. To protect the privacy of private vehicle owners, all sensitive information was removed from the raw trajectories. All researchers were subject to a strict non-disclosure agreement.

### B. Baselines and Metrics

During the detection of regular travelers, we chose classic and advanced algorithms as the baselines.

- **K-means Clustering.** K-means clustering is an unsupervised learning algorithm. We set  $K = 2$ , and the vehicle's category is determined according to the type of label in the clustering result.
- **Density-Based Spatial Clustering of Applications with Noise (DBSCAN).** DBSCAN is a density-based clustering model, in which we set  $eps = 500$  m, and  $minPts = 3$ .
- **Ordering Points To Identify the Clustering Structure (OPTICS).** OPTICS is a density-based model. We set  $MinPts = 3$ , and the range of  $eps$  was [300,1000].
- **Trajectory Aggregation Detection (TAD)** [28]. TAD is an analysis method based on travel features. According to the density of the trajectory data, a dynamic domain adaptive method was used to determine the domain radius and the type of private vehicle.
- **Density-Peak based Clustering (DPC)** [21]. DPC uses the local density of each data and the distance to the nearest neighbor with a higher density to isolate and identify the cluster centers.
- **Extracting regular travel behavior (ERTB)** [33]. ERTB is a transfer learning method that utilizes existing knowledge to identify regular travel behavior from private vehicle trajectory data.

For the validation and comparison of performance between different methods, we used the labeled samples in the sample set as the unlabeled samples, and the remaining 200 samples as the validation set. The three commonly used evaluation indicators in a detection method are accuracy, precision, and recall, all of which have been used for the performance evaluation conducted in our study:

$$ACC = \frac{TP + TN}{N}, \quad (15)$$

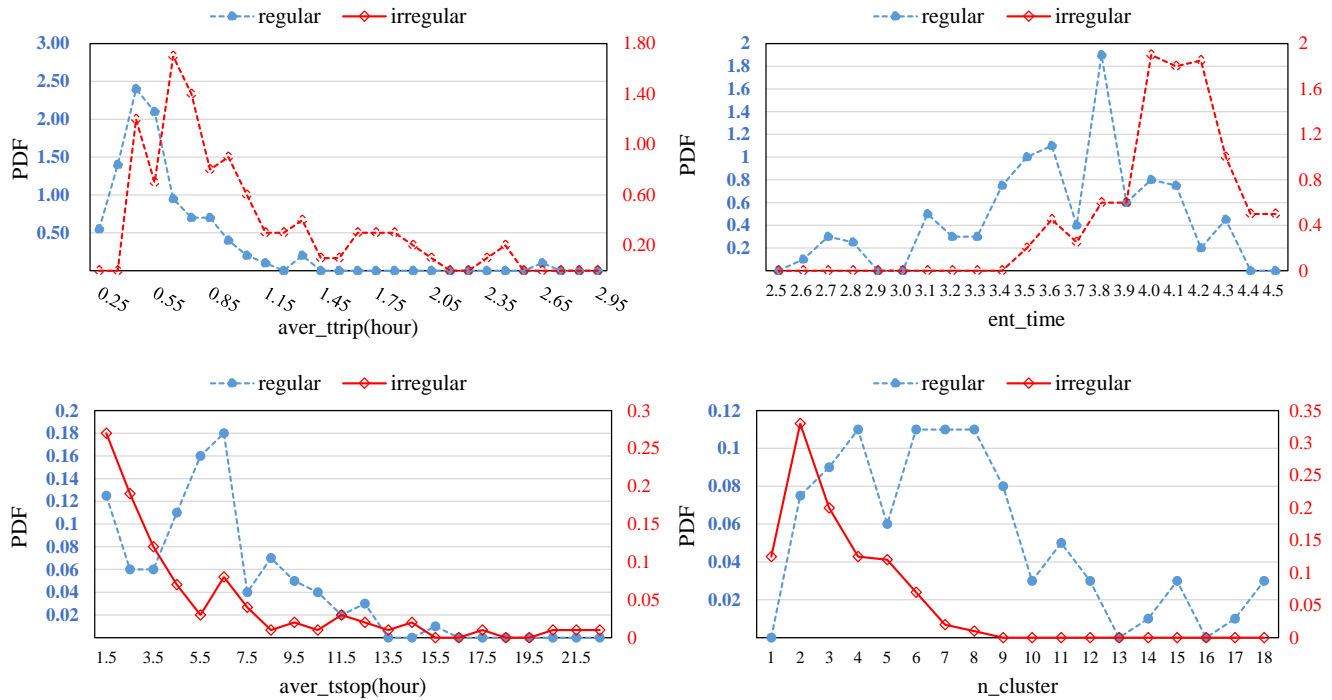


Fig. 7. Feature distribution of travelers. (a) Average trip time. (b) Entropy of the move time. (c) Average stay duration. (d) Frequent clusters.

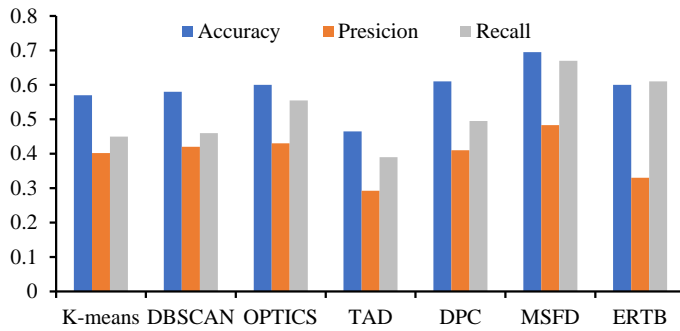


Fig. 8. Comparison results.

$$P = \frac{TP}{TP + FP}, \quad (16)$$

$$R = \frac{TP}{TP + FN}. \quad (17)$$

where  $N$  represents the number of labeled samples in the sample set,  $TP$  represents the number of samples whose labels are correctly predicted as regular travelers,  $TN$  represents the number of samples whose labels are correctly predicted as irregular travelers,  $FP$  represents the number of samples that are labeled incorrectly as irregular travelers, and  $FN$  represents the number of samples that are incorrectly labeled as regular travelers.

### C. Results and Discussion

We use the probability density function (PDF) to further analyze the travel features of regular travelers in the experimental results. As shown in Figure 7(a), the average move duration of regular travelers is primarily within 1 h, whereas irregular

travelers have a longer move duration. Therefore, regular travelers tend to travel for short durations and distances. We analyze the entropy of the move time of private vehicle users. The higher the entropy value, the higher is the mixing degree and time irregularity. As shown in Figure 7(b), compared to regular travelers, irregular travelers have an entropy of move time that has a greater probability of being a high value. As shown in Figure 7(c), regular travelers have the greatest probability of staying in their destination for 5–7 hours, whereas the average stay duration of irregular travelers is mostly approximately 1 h. Figure 7(d) shows that regular travelers have more frequent stay locations than irregular travelers. A possible reason for this is that the locations where regular travelers generally conduct activities are a part of their daily routines. The locations where irregular travelers conduct activities are random, and most irregular travelers only have two or three locations where they frequently stay.

Figure 8 presents the experiment results in the identification of regular travelers using K-means, DBSCAN, OPTICS, TAD, DPC, and MSFD. The results of MSFD show the highest accuracy (0.695%), precision (0.483%), and recall (0.67%) amongst all the baselines that were compared. Compared to DPC, OPTICS, and DBSCAN, which are three density-based models, MSFD shows a better performance than density-based clustering. Specifically, compared to DPC, MSFD improves the accuracy by 13.9%; precision, by 17.8%; and recall, by 35.3%. Compared to OPTICS, MSFD improves the accuracy by 15.8%; precision, by 12.3%; and recall, by 20.7%. Compared to DBSCAN, MSFD improves the accuracy by 20.2%; precision, by 15.0%; and recall, by 45.6%. The aforementioned results can be explained by the fact that supervised learning methods only use labeled training samples. However,

the amount of training data is insufficient to train a better model. Compared to K-means, MSFD improves the accuracy by 21.9%; precision, by 20.2%; and recall, by 48.9%. As a possible reason for this, K-means does not add supervision constraints. The results show that the addition of supervision constraints in our study makes MSFD more practical. Compared to TAD, MSFD improves the accuracy by 49.5%; precision, by 65.2%; and recall, by 71.8%. This is because TAD is based on the aggregation features of the vehicles, whereas MSFD is based on the features of the vehicle travel behavior. Compared to ERTB, MSFD improves the accuracy by 15.8%; precision, by 46.4%; and recall, by 9.8%. The results show that it is more reasonable to have the travel behavior features of the vehicle owners as the basis of the detection.

## VII. CONCLUSIONS

In this paper, we proposed move-and-stay feature based methodologies for performing a regular travel behavior analysis and regular traveler discovery. We extracted the move-and-stay frequency, time, duration, and distance features from a huge private vehicle trajectory dataset to represent the travel behavior. We then proposed a method that leverages the sum of the squared errors to fit the regularities of these extracted features. The analysis results show that all of these features follow different distributions on weekends and weekdays, respectively. Specifically, the move duration and move distance follow an inverse gamma distribution on weekdays and an exponential Weber distribution on weekends. The stay duration follows an inverse Gaussian distribution on weekdays, and on weekends, it follows a generalized Pareto distribution. Moreover, based on the move-and-stay features, we proposed a semi-supervised model to identify regular travelers from massive private vehicle trajectory data. Quantitative experimental results show that the proposed model can achieve higher accuracy in regular traveler identification than other baselines.

Further efforts will be made toward taking a deeper look into regular travel behavior by considering other factors such as weather, holidays, and special events. Moreover, because regular travelers that share common spatiotemporal travel features have similar travel demands, our results can provide an empirical basis for shared mobility services.

## REFERENCES

- [1] Armando Bazzani, Bruno Giorgini, Sandro Rambaldi, Riccardo Gallotti, and Luca Giovannini. Statistical laws in urban mobility from microscopic gps data in the area of florence. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(05):1–17, 2010.
- [2] Irad Ben-Gal, Shahar Weinstock, Gonen Singer, and Nicholas Bambos. Clustering users by their mobility behavioral patterns. *ACM Transactions on Knowledge Discovery from Data*, 13(4):1–28, 2019.
- [3] Meng Chen, Xiaohui Yu, and Yang Liu. Mpe: A mobility pattern embedding model for predicting next locations. *World Wide Web*, 22(6):2901–2920, 2019.
- [4] Carmela Comito, Deborah Falcone, and Domenico Talia. Mining human mobility patterns from social geo-tagged data. *Pervasive and Mobile Computing*, 33:91–107, 2016.
- [5] L Espn-Noboa, F Lemmerich, P. Singer, and M. Strohmaier. Discovering and characterizing mobility patterns in urban spaces: A study of manhattan taxi data. *International World Wide Web Conferences Steering Committee*, pages 1–6, 2016.
- [6] Riccardo Gallotti, Armando Bazzani, Sandro Rambaldi, and Marc Barthelemy. A stochastic model of randomly accelerated walkers for human mobility. *Nature communications*, 7(1):1–7, 2016.
- [7] Gabriel Goulet-Langlois, Haris N Koutsopoulos, Zhan Zhao, and Jinhua Zhao. Measuring regularity of individual travel patterns. *IEEE Transactions on Intelligent Transportation Systems*, 19(5):1583–1592, 2017.
- [8] Md Mehedi Hasnat and Samiul Hasan. Identifying tourists and analyzing spatial patterns of their destinations from location-based social media data. *Transportation Research Part C: Emerging Technologies*, 96:38–54, 2018.
- [9] Vincent Havyarimana, Damien Hanyurwimfura, Philibert Nsengiyumva, and Zhu Xiao. A novel hybrid approach based-srg model for vehicle position prediction in multi-gps outage conditions. *Information Fusion*, 41:1–8, 2018.
- [10] Vincent Havyarimana, Zhu Xiao, Alexis Sibomana, Di Wu, and Jing Bai. A fusion framework based on sparse gaussian-wigner prediction for vehicle localization using GDOP of GPS satellites. *IEEE Transactions on Intelligent Transportation Systems*, 21(2):680–689, 2020.
- [11] Yourong Huang, Zhu Xiao, Dong Wang, Hongbo Jiang, and Di Wu. Exploring individual travel patterns across private car trajectory data. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–15, 2019.
- [12] Yourong Huang, Zhu Xiao, Xiaoyou Yu, Dong Wang, Vincent Havyarimana, and Jing Bai. Road network construction with complex intersections based on sparsely sampled private car trajectory data. *ACM Transactions on Knowledge Discovery from Data*, 13(3):1–28, 2019.
- [13] Hongbo Jiang, Yu Zhang, Zhu Xiao, Ping Zhao, and Arun K Iyengar. An empirical study of travel behavior using private car trajectory data. *IEEE Transactions on Network Science and Engineering*, 8(1):53–64, 2020.
- [14] Kyoungok Kim. Identifying the structure of cities by clustering using a new similarity measure based on smart card data. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):2002–2011, 2019.
- [15] Dániel Kondor, Hongmou Zhang, Remi Tachet, Paolo Santi, and Carlo Ratti. Estimating savings in parking demand using shared vehicles for home-work commuting. *IEEE Transactions on Intelligent Transportation Systems*, 20(8):2903–2912, 2018.
- [16] Dheeraj Kumar, Huayu Wu, Sutharshan Rajasegarar, Christopher Leckie, Shonali Krishnaswamy, and Marimuthu Palaniswami. Fast and scalable big data trajectory clustering for understanding urban mobility. *IEEE Transactions on Intelligent Transportation Systems*, 19(11):3709–3722, 2018.
- [17] Leonidas Mantzos, Pantelis Capros, Nikolaos Kouvaritakis, and M Zeka-Paschou. European energy and transport-trends to 2030. 2009.
- [18] K Mohamed, Etienne Côme, Latifa Oukhellou, and Michel Verleysen. Clustering smart card data for urban mobility analysis. *IEEE Transactions on intelligent transportation systems*, 18(3):712–728, 2016.
- [19] Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Albert-László Barabási. Returners and explorers dichotomy in human mobility. *Nature communications*, 6(1):1–8, 2015.
- [20] Geqi Qi, Ailing Huang, Wei Guan, and Lingling Fan. Analysis and prediction of regional mobility patterns of bus travellers using smart card data and points of interest data. *IEEE Transactions on Intelligent Transportation Systems*, 20(4):1197–1214, 2018.
- [21] Alex Rodriguez and Alessandro Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.
- [22] Iqbal H Sarker, Alan Colman, Muhammad Ashad Kabir, and Jun Han. Individualized time-series segmentation for mining mobile phone user behavior. *The Computer Journal*, 61(3):349–368, 2018.
- [23] Zhangqing Shan, Weiwei Sun, and Baihua Zheng. Extract human mobility patterns powered by city semantic diagram. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–14, 2020.
- [24] Liridona Sopjani, Jenny Janhager Stier, Mia Hesselgren, and Sofia Ritzén. Shared mobility services versus private car: Implications of changes in everyday life. *Journal of Cleaner Production*, 259:120845–120857, 2020.
- [25] Etienne Thuillier, Laurent Moalic, Sid Lamrous, and Alexandre Caminada. Clustering weekly patterns of human mobility through mobile phone data. *IEEE Transactions on Mobile Computing*, 17(4):817–830, 2017.
- [26] Chengcheng Wan, Yanmin Zhu, Jiadi Yu, and Yanyan Shen. Smopat: Mining semantic mobility patterns from trajectories of private vehicles. *Information Sciences*, 429:12–25, 2018.
- [27] Dong Wang, Jiaojiao Fan, Zhu Xiao, Hongbo Jiang, Hongyang Chen, Fanzi Zeng, and Keqin Li. Stop-and-wait: Discover aggregation effect based on private car trajectory data. *IEEE transactions on intelligent transportation systems*, 20(10):3623–3633, 2018.



- [28] Dong Wang, Qian Liu, Zhu Xiao, Jie Chen, Yourong Huang, and Weiwei Chen. Understanding travel behavior of private cars via trajectory big data analysis in urban environments. In *IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing*, pages 917–924, 2017.
- [29] Feilong Wang and Cynthia Chen. On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies*, 87:58–74, 2018.
- [30] Feng Xia, Jinzhong Wang, Xiangjie Kong, Zhibo Wang, Jianxin Li, and Chengfei Liu. Exploring human mobility patterns in urban scenarios: A trajectory data perspective. *IEEE Communications Magazine*, 56(3):142–149, 2018.
- [31] Zhu Xiao, Fancheng Li, Ronghui Wu, Hongbo Jiang, Yupeng Hu, Ju Ren, Chenglin Cai, and Arun Iyengar. Trajdata: On vehicle trajectory collection with commodity plug-and-play obu devices. *IEEE Internet of Things Journal*, 7(9):9066–9079, 2020.
- [32] Zhu Xiao, Pingting Li, Vincent Havyarimana, Georges Maigary Hassana, Dong Wang, and Keqin Li. Goi: A novel design for vehicle positioning and trajectory prediction under urban environments. *IEEE Sensors Journal*, 18(13):5586–5594, 2018.
- [33] Zhu Xiao, Shenyan Xu, Tao Li, Rui Zhang, Hongbo Jiang, Amelia Regan, and Hongyang Chen. On extracting regular travel behavior of private cars based on trajectory data analysis. *IEEE Transactions on Vehicular Technology*, 69(12):14537– 4549, 2020.
- [34] Rong Xie, Yang Chen, Qinge Xie, Yu Xiao, and Xin Wang. We know your preferences in new cities: Mining and modeling the behavior of travelers. *IEEE Communications Magazine*, 56(11):28–35, 2018.
- [35] Junjun Yin, Yizhao Gao, Zhenhong Du, and Shaowen Wang. Exploring multi-scale spatiotemporal twitter user mobility patterns with a visual-analytics approach. *ISPRS International Journal of Geo-Information*, 5(10):187–206, 2016.
- [36] Faming Zhang, Xinyan Zhu, Wei Guo, Xinyue Ye, Tao Hu, and Liang Huang. Analyzing urban human mobility patterns through a thematic model at a finer scale. *ISPRS International Journal of Geo-Information*, 5(6):78–95, 2016.
- [37] Juanjuan Zhao, Qiang Qu, Fan Zhang, Chengzhong Xu, and Siyuan Liu. Spatio-temporal analysis of passenger travel patterns in massive smart card data. *IEEE Transactions on Intelligent Transportation Systems*, 18(11):3135–3146, 2017.
- [38] Ziliang Zhao, Shih-Lung Shaw, Yang Xu, Feng Lu, Jie Chen, and Yin Yin. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, 30(9):1738–1762, 2016.
- [39] Linjiang Zheng, Dong Xia, Li Chen, and Dihua Sun. Understanding citywide resident mobility using big data of electronic registration identification of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 21(10):4363–4377, 2020.