Hindawi
Journal of Advanced Transportation
Volume 2017, Article ID 1738085, 14 pages
https://doi.org/10.1155/2017/1738085

Research Article

Clustering Vehicle Temporal and Spatial Travel Behavior Using License Plate Recognition Data

Huiyu Chen, Chao Yang, and Xiangdong Xu

Key Laboratory of Road and Traffic Engineering, Tongji University, 4800 Cao'an Road, Shanghai 201804, China

Correspondence should be addressed to Chao Yang; tongjiyc@tongji.edu.cn

Received 27 December 2016; Revised 13 March 2017; Accepted 2 April 2017; Published 24 April 2017

Academic Editor: Guohui Zhang

Understanding travel patterns of vehicle can support the planning and design of better services. In addition, vehicle clustering can improve management efficiency through more targeted access to groups of interest and facilitate planning by more specific survey design. This paper clustered 854,712 vehicles in a week using K-means clustering algorithm based on license plate recognition (LPR) data obtained in Shenzhen, China. Firstly, several travel characteristics related to temporal and spatial variability and activity patterns are used to identify homogeneous clusters. Then, Davies-Bouldin index (DBI) and Silhouette Coefficient (SC) are applied to capture the optimal number of groups and, consequently, six groups are classified in weekdays and three groups are sorted in weekends, including commuting vehicles and some other occasional leisure travel vehicles. Moreover, a detailed analysis of the characteristics of each group in terms of spatial travel patterns and temporal changes are presented. This study highlights the possibility of applying LPR data for discovering the underlying factor in vehicle travel patterns and examining the characteristic of some groups specifically.

1. Introduction

The trip starting and ending time, travel distance, travel frequency, activity duration, and some analogous features are the typical form of vehicle travel behaviors. All these aspects have a significant effect on the traffic condition in a direct or indirect way [1, 2]. For example, the distribution of the trip starting and ending time of all vehicles will decide the peak-hour time. Better understanding of these characteristics will be helpful to analyze the travel pattern and travel mode of vehicles. Identifying homogeneous travel behavior groups has been the research subject in several prior studies and the travel behavior analysis has always attracted great interest of transport authorities, since vehicle travel behavior has a vital impact on strategic and operational decisions [3–5].

Clustering is one of the most important methods to count and mine meaningful information in large amount of data since understanding the main differences between groups can contribute to a better understanding of their travel behaviors, which can provide valuable information for transportation planning [6]. Meanwhile, clustering vehicles based on their travel characteristics is one of the vital methods for studying the representativeness of specific groups among the whole vehicle population and the travel profile of each group provides an aggregated characterization for the vehicles of a group as a whole [7]. It can also provide transportation planners with richer travel demand information for improving the system performance or better assessing network investments.

In the field of transportation, clustering has been widely accepted in dealing with big data and traffic problems [8, 9]. Reference [10] investigated the determination of historical traffic patterns by means of Ward's hierarchical clustering procedure. It classifies the traffic patterns in highways with the data collected by automatic vehicle identification (AVI) system into four groups and the resultant weekday traffic patterns can be used as input for macroscopic traffic models and as a basis for traffic management. Moreover, when predicting traffic flows based on historical data, a preclassification (e.g., holidays, Mondays, core weekdays, and Fridays) can be made to guide the authorities, and these patterns can be used to detect and replace erroneous data and to impute missing data.

Besides, [11, 12] utilized the density-based clustering algorithms to classify trajectories using GPS data. The study

of trajectory data can reveal individual trajectory patterns, understand the characteristics of human dynamics, and thus support trajectory prediction, urban planning, traffic monitoring, and so forth. The characteristics will be similar inside each group and significantly different outside the groups. According to the similarity, the individual similar trajectory recognition can be achieved; and by clustering, the abnormal trajectory mode detection can also be conducted. Similarly, literature [13, 14] combined DBSCAN and SVM (support vector machines) cluster algorithm to sort the GPS trajectories to identify the activity stop locations, which has significance in analyzing human urban mobility.

In the analysis of the time series characteristics of traffic flow data [15], clustering method is popular too. According to the similarity of traffic flow characteristics, the traffic sections are divided into different groups and in the literature [16]; performance of the proposed approach and the stability of the clustering technique are evaluated using the extensive simulation for different traffic densities.

Numerous researches concerning traffic and travel have been conducted by previous studies but there are some drawbacks at the same time. It is difficult to obtain the large amount of data. The acquisition is mostly based on artificial method but at the expense of consuming lots of manpower and resources. Worse still, there are much error and abnormities in the information usually; thus, the research results always show lower reliability and higher deviation.

Recently, a number of well-established technologies for collecting vehicle related data have emerged, including loop detectors, GPS data, and probe car data [17, 18]. Loop detectors have the merit that once they are installed, there will be continuous record when every vehicle is passing the monitored road section. However, the share of segments in the network equipped with these sensors is typically low and cannot represent the urban network as a whole, which will leave the traffic conditions in most of the network unknown. Dedicated probe vehicles, meanwhile, are used to collect the travel time and other data for designated routes in the network. Nevertheless, due to cost considerations, the number of traffic studies with probe vehicles is typically small and the number of vehicles involved is very few. Hence, they can only cover a limited number of routes for a limited duration of time.

A number of limitations mean that new sophisticated methods are needed to process the data and generate useful information, compared to traditional sensors [19]. Most recently, with the emerging technologies and advanced devices, image recognition technology has been greatly improved. License plate recognition (LPR) system provides the opportunity to study in detail vehicle travel patterns. Compared to manual data collection techniques, LPR provides lower marginal costs, more detailed and disaggregated information, large sample size, and real-time data availability [20, 21]. LPR data is mainly applied in LPR data is mainly applied in solving three kinds of problems in the field of transportation, that is, (1) road network state discrimination, (2) vehicle microscopic characteristics mining, and (3) vehicle travel time/path estimation [22, 23]. Zhan et al. [24] proposed a lane-based real-time queue length estimation model applying the LPR data. By using ground truth information of the maximum queue length from the city of Langfang in China, the model is validated. In addition, a novel trip route estimation method was given by researchers to estimate the vehicle travel path [25]. Similarly, based on LPR data, an approach for forecasting urban short-term OD matrix which can be used to obtain the original OD information was came up with, and then the OD amount between the detection points can be inferred and finally the OD information between fast track ramps is obtained [26, 27]. All of that mentioned above has proved that the massive amount of LPR data has been created and provides us with rich information and thus can be an effective analytical data source.

Methods for clustering are usually divided into two categories, supervised and unsupervised. Supervised methods use the past data as training samples or previously known outputs to create and learn a clustering rule that allows the clustering of future or new observations [28]. Because the form of the data is not fitting for this study, unsupervised methods are more applicable. Unsupervised cluster algorithms include the hierarchical algorithms and the partition algorithms. Hierarchical clustering algorithms have high computational complexity and cost, limiting their application to large-scale data sets and the shortcomings and advantages of these algorithms will be explained in the following paragraph.

K-means clustering algorithm, which belongs to the distance-based clustering algorithms, is not only the most classic, but also the most widely used. It has the property of rapid computing speed, easily explained principle, and high efficiency. *K*-means clustering algorithm is tested using load profiles of 100 residential smart meters collected over the interval extending from July 20th until August 9th, 2009. The method has shown high accuracy in dealing with traffic problems, which proved its great applicability [29].

In this paper, data from LPR system in Shenzhen, China, from November 4th to 10th, 2013, during seven days (a week) in total are analyzed. Variables chosen for clustering include the proportion of different starting/ending points, maximum/minimum/average travel distance for one trip, days of travel within a week, the number of trips per day, the average start time of the first trip, the average end time of the last trip, and activity duration [30]. Firstly, data cleaning is conducted to remove the wrong and repeated data. Then, deviation standardization is utilized to normalize each value for eliminating the error caused by dimension and considerable differences of magnitude. After preliminary treatment, data is divided into two groups, namely, the weekdays and weekends. Finally, to measure the optimal number of clusters, Davies-Bouldin index (DBI) [31] and Silhouette Coefficient (SC) [32] are employed.

In general, the purpose of this study is to classify vehicles into several categories based on some variables and determine travel behavior consistency over time and space by analyzing the vehicle temporal and spatial variability. It can support the study of representing specific groups among the

Figure 1: Road network and distribution of LPR system in Shenzhen.

Table 1: Raw data sample.

Vehicle ID	Detector ID	Lane number	Date and time
2a0adafb3bedf3ad09730c47e0b195b5	20400801	3	2013/11/04 23:00
686ded4bef182aeb03cf361eb6ac6f65	101A0753	2	2013/11/05 00:13
88765784c6cd4464ded7f2336c24eaf2	20602701	1	2013/11/06 14:26
14c397a9bc79e52854194e6c49a69b3e	30606705	2	2013/11/07 15:03
6c43c9bb61e730194de304f0ce9e99ae	206A0480	1	2013/11/08 09:21
fb567b73e8e3db9fa03a4d265b6bddbe	20605302	2	2013/11/09 04:18
45e8dca9594b91c5fdb40706b6f0abc7	10100210	3	2013/11/10 19:14

total population and help establish the predictive level of vehicle trips.

The rest of the paper is organized in the following way. Section 2 offers a brief description of data source. The methodology is introduced in Section 3. Section 4 displays the variables chosen for clustering. Section 5 shows the results of the clustered data and Section 6 is the conclusion and findings.

2. Data Description

2.1. Data Overview. The potential of LPR system has been explored for planning, managing, and assessing the performance of traffic systems. Further, data collected by these systems allows more comprehensive view of vehicle travel patterns and travel behaviors.

2.1.1. Data Source. The LPR system in Shenzhen, China, covers majority of parking lots and expressways for this city. Over 0.9 million vehicles are detected in a week and according to Shenzhen Statistical Yearbook in 2013, the total number of vehicles in Shenzhen is about 2.1 million, implying 42.86% of vehicles are detected by the LPR system. After data cleaning, there are still almost 128,000 recorded vehicles each day. Figure 1 is the sketched network of Shenzhen, where the red points represent the detectors installed on roads and the black lines show the roads.

LPR detectors are mainly installed in the expressways of the city unevenly, most of which are on the intersection or the pedestrian bridge nearby. They are denser in the city center area, while more are dispersed in the rest of the region. The sample of raw data is given in Table 1.

It is worth noting that the detector ID has two types, 10100610 and 101A0753. If "A" is contained in the ID, the detector is a parking lot. Otherwise, it represents a detector on road. Table 2 shows the amount of detectors for each day from November 4th to November 10th, 2013, for which more than 83% detectors are parking lots.

Table 2: Amount of detection ID for each day.

Date	Nov. 4	Nov. 5	Nov. 6	Nov. 7	Nov. 8	Nov. 9	Nov. 10
All detectors	918	942	936	934	933	910	873
Parking lots	759	783	777	775	774	751	717
On roads	159	159	159	159	159	159	156

There are three main types of parking lots, (1) residential parking lots (including residential and office buildings, commercial places, and shared parking lots), (2) temporary parking lots, and (3) public parking lots. The parking lots with detection data account for about 20% of all the parking lots in Shenzhen.

2.1.2. Data Cleaning. The data cleaning is conducted before vehicle clustering and there are two main steps.

(1) Extract the Data by Day. The whole dataset is for seven days (a week), which has been separated into seven files by date thus each file contains the data of the same day.

(2) Verify the Original LPR Data

(1) Delete erroneous LPR data: there are two kinds of erroneous data in our study: (a) the detected time of the record is beyond the range of [0:00–24:00] and (b) the latitude and longitude of the detection site of the record are beyond the scope of Shenzhen.

(2) Remove duplicated LPR data records: if there are two identical records, only one needs to be kept.

(3) Extract the trip chain in accordance with the definition of one trip: that is, the data has been processed into the following form.

Vehicle a-time(1)- location(1), vehicle a-time(2)- location (2),. . .,vehicle a-time(n)- location (n)
Vehicle b-time(1)- location (1), vehicle b-time(2)- location (2),. . .,vehicle b-time(n)- location (n)

Based on the trip chain of the vehicles, all of the mentioned variables can be calculated, such as the trip starting time, ending time, the whole activity duration, and the travel distance.

2.2. Identification of Taxi. The purpose of this paper is to cluster all the vehicles in the dataset according to some temporal and spatial variables. Each group will have some characteristics different from the other groups, so as to explore vehicle travel patterns we may not know before. Traffic researchers have always paid much attention to taxi, due to its special travel mode. It has the following characteristics [33]:

(1) There are no fixed route and running time.

(2) Operation is for 24 hours and can be located in any place of the city.

(3) The origins and destinations of taxi are completely determined by passengers.

(4) The operating routes are up to the driver, such as his experience and hobbies.

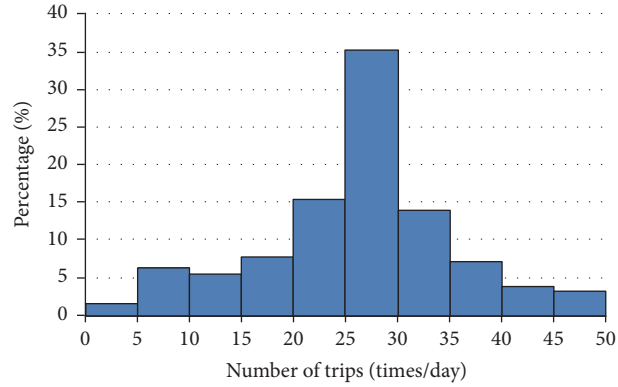

Figure 2: The distribution of number of trips for taxi in Shenzhen.

On account of these features, taxis are removed from the dataset to make sure the analysis in this paper is more specific on noncommercial vehicles and the future research will focus on the travel behavior of taxi.

Figure 2 shows the distribution of the number of taxi trips in Shenzhen. There are over 70% of vehicles traveling 20–35 trips per day, and only 7% vehicles are traveling less than 10 trips. Meanwhile, from the clustering result, the travel frequency of nontaxis in a day is no more than 10 trips per day. As a result, we removed vehicles whose travel frequency exceeded 10 trips per day. Under such a definition, there may be two inaccurate results: (1) nontaxis traveling more than 10 times a day were removed and (2) taxis traveling less than 10 times were still retained.

However, in the light of Shenzhen Statistical Yearbook in 2013, the number of taxis is around 17,000 in total, in which less than 50% were detected by the LPR system. Thus, the amount of these two kinds of vehicles will be no more than a thousand, which appears insignificant when compared with tens of thousands ordinary vehicles.

On the basis of the rule proposed above, almost 6,000 taxis for one day are removed from the dataset and when taxis are removed, there are around 122,000 vehicles for each day and 854,000 vehicles in a week.

3. Methodology

3.1. Clustering Methods. Clustering methods encompass several techniques and algorithms used to group observations based on similar qualitative or quantitative characteristics. They are usually divided into supervised and unsupervised clustering. Supervised methods require a training sample which contains previously known information on each group membership [34]. In accordance with the form of data in this

Table 3: Advantages and disadvantages of some unsupervised algorithms.

Clustering algorithm	Representational algorithm	Advantage	Disadvantage
Hierarchical algorithms	BIRCH CURE Chameleon	(1) No input parameters are required (2) High scalability	(1) high computational complexity and cost; (2) low efficiency in dealing with large-scale data
Partition algorithms	K-means CLARANS	(1) High efficiency in dealing with large-scale data (2) Fast calculation speed	(1) dependent on initial center selection; (2) uncertainty of category number

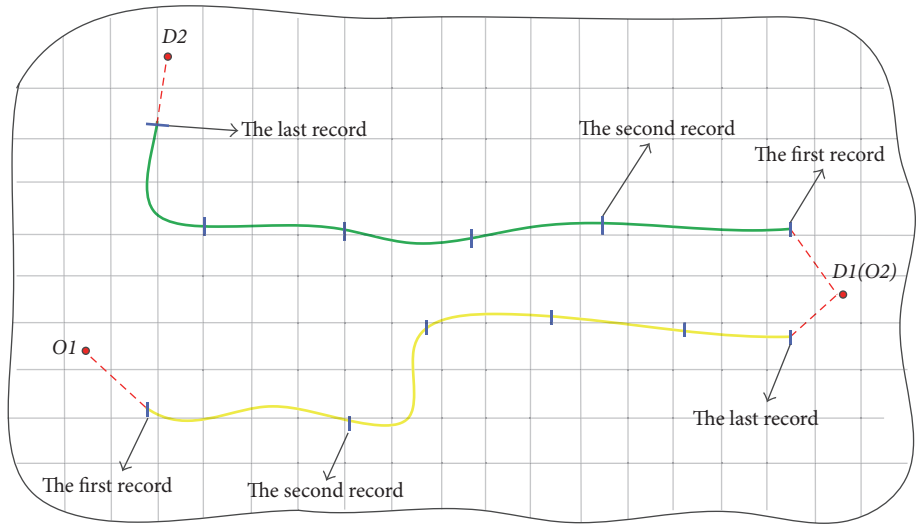

Figure 3: The relationship between the travel trajectory and the detection points.

study, the training sample is not available and there are no previously known classes; unsupervised clustering method is the best option. Unsupervised clustering methods aim at categorizing the data objects without a training sample; the goal is to find clusters based on similarities of the input data. There are two main types of unsupervised clustering, the hierarchical algorithms and the partition algorithms. Table 3 discusses the advantages and disadvantages of some unsupervised algorithms [35].

3.2. K-Means Algorithm. As shown in Table 3, hierarchical algorithms have been criticized for low robustness and high sensitivity to noise and outliers. Since the assignment of an object to a cluster is not iterative, hierarchical algorithms are not able to correct potential misclassifications. On the contrast, partition algorithms optimize either a locally or a globally defined objective function to generate groups of observations so they are preferred in studies involving large-scale dataset.

K-means is chosen for this study as a computationally efficient method, which is suitable for situations where all variables are quantitative. It is easy to understand and apply and thus is popular in dealing with the clustering problems. The time complexity of K-means algorithm is close to linear, is simultaneously suitable for mining large-scale data sets, and is scalable. In this study, the variables used for clustering are all quantitative and we have a large amount of data. So, K-means is chosen for this study. Nevertheless, the only disadvantage is the difficulty of choosing the number of clusters and their dependency on the initialization scenario. For the first drawback, it can be adjusted by repeated iterations to find the optimal result. For the second one, we have tried several cluster numbers and applied Davies-Bouldin index (DBI) and Silhouette Coefficient (SC) to find the optimal cluster number.

3.3. Criteria for One Trip. For the sake of turning the raw data into the form of vehicle trips and the value of its corresponding variables, the criteria for one trip should be given firstly. Due to the inherent limitation of the LPR data, only partial trajectory points of a vehicle can be obtained. As a result, the realistic starting and ending points of a trip cannot be speculated.

Figure 3 shows the travel trajectory of a vehicle in a brief network, where the yellow curve represents the first trip of the vehicle and the green one displays its second trip. In addition, the blue short lines show the detecting points. It is definite that the true trip starting time in origin 1 ($O1$) is earlier than the time of the first trip record, and the trip ending time in destination 1 ($D1$) is later than the time of the last record, as well as the second trip or other trips of the vehicle.

Hence, deviation will exist in the value of some variables inevitably. The average starting time of the first trip will be a little later, and the average ending time of the last trip will be

Table 4: The average number of trips for different thresholds.

Threshold (min)	Number of trips	Threshold (min)	Number of trips
20	4.52	50	1.96
25	3.86	55	**1.83**
30	3.24	60	**1.82**
35	2.90	65	**1.82**
40	2.41	70	**1.81**
45	2.13	75	1.78
50	1.96	80	1.73

a little earlier. The whole activity duration will be longer and the travel distance will be shorter. However, the main purpose of our study is to extract the travel characteristics of vehicles instead of the estimation of the *OD* matrix; these errors are offset in one direction for all vehicles; thus it may not have a critical impact on the clustering result. From this point of view, the definition of one trip is applicable. When applying these values of variables in realistic transportation planning, the deviation should be taken into account.

As mentioned, the "segmentation" refers to the interval between two trips that is the interval of the last record of the first trip and the first record of the second trip, which is different from the vehicle's accumulated travel time. In order to find the optimal value of the segmentation, we have tested the threshold.

Set AR to be the true threshold, BR to be the true number of trips, A to be the threshold that we will apply, and B to be the number of trips that we will calculate. If $A \leq \text{AR}$, then $B \geq \text{BR}$; if $A \geq \text{AR}$, then $B \leq \text{BR}$; only when $A = \text{AR}$, then $B = \text{BR}$. Different thresholds ranging from 20 min to 80 min have been tested, and the average number of trips under all circumstance is calculated. The result was illustrated as Table 4.

When the threshold spans from 50 min to 80 min, the value of number of trips has been moving towards stabilization. It implies that the probability of trips to be not detected in this interval is relatively small. Also, the interval of two trips from LPR data is larger than the actual interval. Hence, it is reasonable that one hour is chosen to be the threshold.

4. Clustering Variables

4.1. Spatial and Temporal Variables. To estimate homogeneous vehicle groups based on their travel patterns using any clustering method, it is necessary to have input information on travel behaviors. Travel patterns can be described by looking at specific variables that together characterize each vehicle's travel routines [36]. The selected variables must include those vehicles' characteristics that make their travel patterns distinct [37, 38]. A set of descriptive variables is presented and vehicles are analyzed in weekdays and weekend separately.

(1) The Proportion of Different Origins/Destinations. The percentage of different origins/destinations has the potential to be a useful indicator of their mobility patterns. To illustrate, vehicles with the same starting point for the first trip in a day or the same ending point for the last trip in a day over a week are more likely to be commuters with work or study purposes. This variable is an indicator of spatial travel variability, which could help to infer the vehicle travel predictability. For such vehicles that traveled 3 days in weekdays, the percentage of different origins for the first trip in a day is defined as follows:

0: The origins of the first trip in a day over the three days are all the same.

1/3: There is one difference for the origins of the first trip in a day over the three days.

2/3: There are two differences for the origins of the first trip in a day over the three days.

1: The origins of the first trip in a day over the three days are all different.

When the value is 0, the origins for one trip are all the same in the days of travel, suggesting that the behavior of this kind of vehicles has much regularity. In contrast, if the value is 1, the origins for one trip are all different in the days of travel, indicating the irregularity of the travel behaviors.

The calculation for percentage of different destinations for the last trip in a day is defined in the same way, and for vehicles in weekends the dealing method is comparable.

(2) Travel Distance. The geometric distance between the origin and destination of one trip can show how accessible activity locations are to a vehicle. Travel distance variability among the trip of a vehicle can also demonstrate travel flexibility and vehicle mobility around the city. The travel distance variables adopted in this study incorporate the maximum/minimum/average travel distance for one trip in the whole week. For the lack of the track points, complete travel trajectory of one trip for a vehicle cannot be obtained. As a result, in this study, the distance of one trip for a vehicle is defined as the exact distance between the start and end points of one trip, which is calculated by the latitude and longitude of the two points.

(3) Travel Frequency. The travel frequency of vehicles, that is, trips made over a day/a week (or any other period) incarnates the uncertainty of the travel for vehicles. There are two descriptive variables, number of trips per day, which is the number of complete trips performed on each day of the week and days of travel, which is the number of days within the

period of analysis; a vehicle has at least one trip in a day. For vehicles in weekdays and weekends, the value of their travel days in a week ranges from zero to seven.

(4) The Trip Start/Finish Time. The trip start/finish time could give expression to the trip purpose and consistency of trip. Volatility of the start time for the first trip and the finish time for the last trip are crucial aspects when analyzing vehicle travel patterns.

(5) Total Activity Duration. Activity refers to all those actions vehicles perform when they are not traveling and in this paper the time interval between the two adjacent trips is defined as the protocol of activity duration. There is a mass of activities purposes, business, work, study, and entertainment, among others. The characteristics of the activity performed at a destination may determine the vehicle's travel decision and the average activity duration of a vehicle in each day varies from weekdays to weekends.

4.2. The Distribution of the Variables for All Vehicles

(1) Weekdays. Figure 4 illustrates the distribution of all the temporal and spatial variables in weekdays which is a statistical indicator of the whole vehicles.

In Figure 4(a), there is an obvious peak during the interval of 8:30 am to 9:00 am, representing that the average trip start time of vehicles is mostly focused between *8:30 am* and *9:00 am*, implying the morning peak hours. Figure 4(b) shows the tendency of the average trip finish time and the majority of the vehicles finish their trip at around *18:00 pm–19:30 pm*, which means the afternoon peak hour. Additionally, there is also a large amount of vehicles that start their trip at 12:30 pm–13:30 pm.

For the number of trips per day in Figure 4(c), vehicles traveling *1.5* trips/day occupy a high proportion and vehicles traveling 3.5 trips/day, 2 trips/day, and 4 trips/day followed. The result seems to be confused that vehicles traveling 1.5 trips/day (less than 2 trips/day) conquer such a high rate. Probably, it is because the definition of one trip in the study and the incomplete vehicle detection data.

Figure 4(d) demonstrates days of travel. Vehicles that only travel one day in a week occupy a high rate. The activity duration of most vehicles is within *11 h* in Figure 4(e). Figures 4(f), 4(g), and 4(h) reflect the travel distance of vehicles. The maximum travel distance of vehicles for one trip is almost within 60 km, the minimum travel distance is less than 30 km, and the average travel distance is within *40 km*. At the same time, we can see that, for the average travel distance of vehicles for one trip, over 68% of trips are within *10 km*.

According to Figures 4(i) and 4(j), for the percentage of different starting or ending points, values 0 and 1 seize on a high proportion. Value 0 means the starting/ending points of each trip are identical, and the regularity is high. Analogously, value 1 means that the starting/ending points of each trip are all different, and irregularity is high.

(2) Weekends. For vehicles traveling in weekends, the distribution of their temporal indicators is basically similar to the weekdays. For the value of both of the percentages for different starting and ending points in weekends, value 0 takes up the highest ratio; in other words, these vehicles travel with less regularity. Compared with the weekday vehicles, they travel a relatively short distance; whether it is the maximum travel distance, minimum travel distance, or average travel distance, almost all are within 10 km and relatively concentrated within 5 km.

5. Results and Discussions

The values of within-cluster variation and the DBI/SC are shown as functions of the number of clusters in Figures 5(a) and 5(b). A smaller value of DBI and a larger value of SC are better. In Figure 5(a), when the cluster number is six, the value of DBI is the smallest, and when it turns to seven, the value of SC is the largest. The value of SC of seven groups is just a little better than six groups but the value of DBI of six groups is much better than seven groups. As a result, "six" is a relatively better choice. In Figure 5(b) when the cluster number is three, both values of SC and DBI are optimal; there is a lowest point of DBI and a highest point of SC. So, the cluster number for weekdays and weekends is selected as six and three, respectively. The K-means clustering method provides not only information about each cluster's core characteristics but also information about the average characteristics of each cluster. Tables 5 and 6 display the average values of each index for each category in weekdays and weekends.

For Vehicles in Weekdays, Six Groups Are Clustered. The last column of Table 5 illustrates the proportion of the total number of each category. The smallest cluster contains 4.1% of the vehicles in the sample, and the largest one accounts for 33.7%. Groups 1 to 6 are identified as follows, long travel distance vehicles, commuting vehicles, noon travel vehicles with short travel distance, off-peak hour travel vehicles, midnight travel vehicles, and peak-hour travel with short activity duration vehicles, respectively.

Group 1 is inferred as long travel distance vehicle that travels 1.82 days in a week and makes 2.13 daily trips. On average, the first trip starting time of Group 1 is 10:14 am and the last trip ending time is 19:02 pm. Additionally, the travel behavior of this group is irregular because the trip origins and destinations are all different. Besides, the total activity duration of this group is about 7.41 hours, and the travel distance of this group of vehicles is relatively long. The maximum travel distance for one trip is 78.1 km on average.

Group 2 may be commuting vehicle, which travels 5.94 days of the week on average and makes 2.18 trips per day. The first trip of the day starts at approximately 8:42 am and the last trip of the day ends at 18:18 pm. The activity duration lasts 8.67 hours on average. Furthermore, the distance between the origin and destination of their trips varies from 6.9 km to 59.2 km, and their average travel distance is about 17.5 km for one trip. The proportion of different starting and ending points for Group 2 is 0.12 and 0.09, representing a high regularity in the daily origins and destinations. All of these features support the speculation of Group 2 to be commuting vehicles.

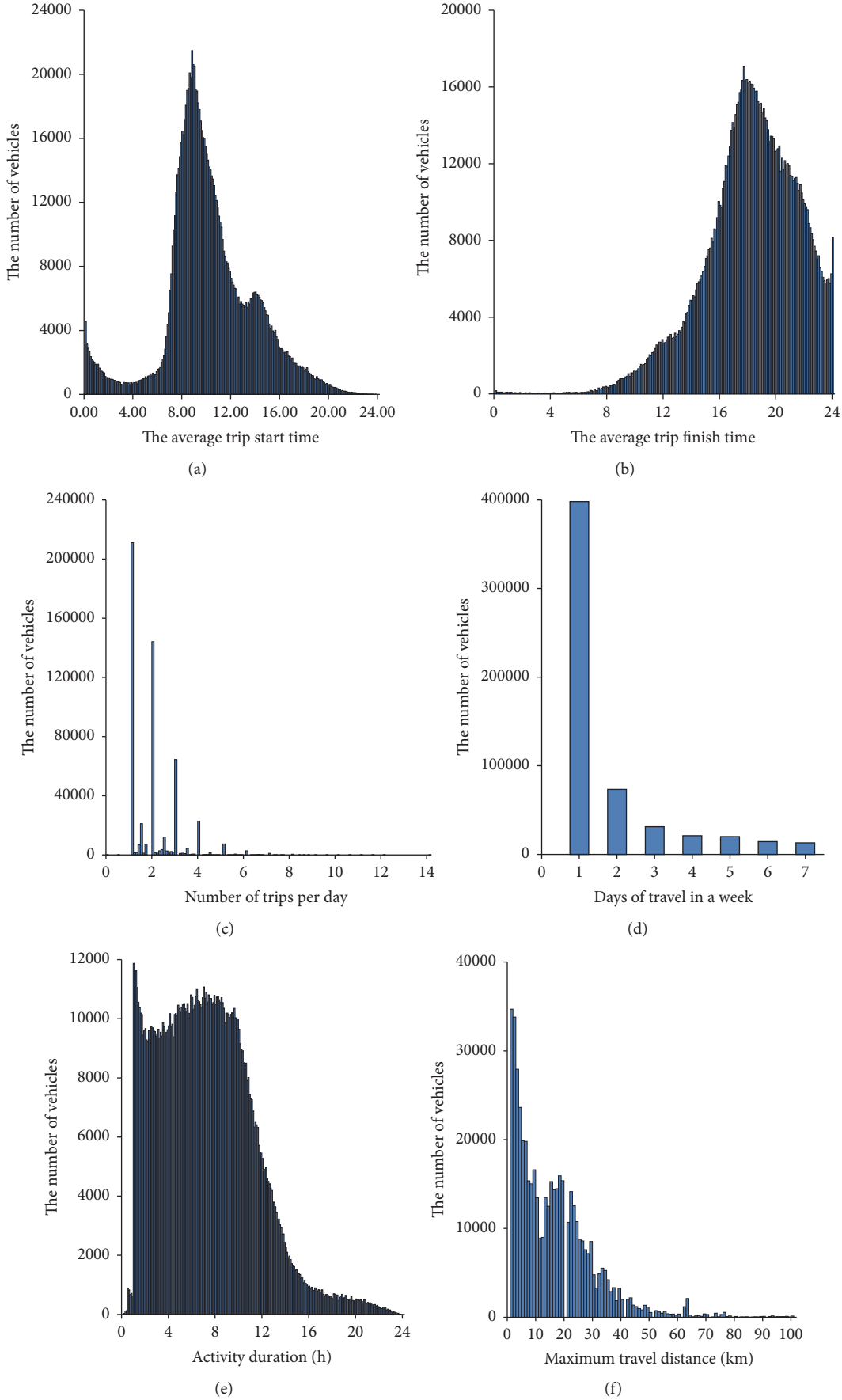

Figure 4: Continued.

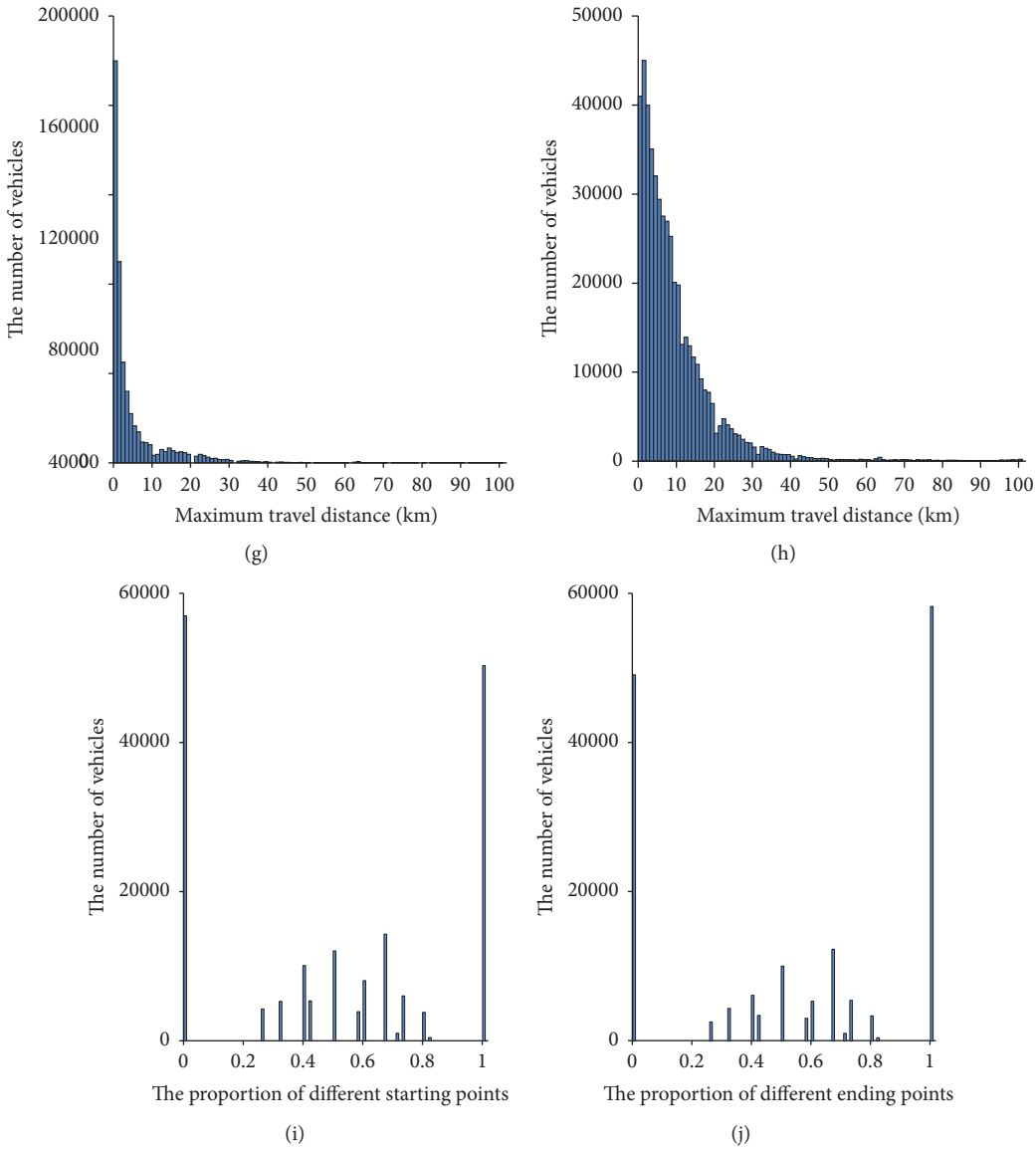

Figure 4: The distribution of variables in weekdays. (a) The average trip start time. (b) The average trip finish time. (c) Number of trips per day. (d) Days of travel. (e) Total activity duration. (f) Maximum travel distance. (g) Minimum travel distance. (h) Average travel distance. (i) The proportion of different staring points. (j) The proportion of different ending points.

Group 3 is defined as noon travel vehicle with short travel distance. The first travel starts at 10:07 am and the last travel ends at 15:14 pm; it only travels at noon. Moreover, Group 3 travels only 1.08 days in a week and 1.82 trips in a day, and the activity duration is also short, only 4.02 hours on average. The travel distance varies between 1.9 km and 3.5 km, dropping in a short range and the travel origins and destinations are almost different.

Group 4 is concluded to be off-peak hour travel vehicle; the first trip of the day starts at 10:20 am and the last trip of the day ends at 19:48 pm, which staggers the peak hours. There are 1.82 days of travel in a week and 1.63 trips in a day and the travel distance of Group 4 is similar to that of Group 3. In particular, the maximum travel distance is only 2.9 km and in accordance with the percentage of different starting and ending points, the travel for Group 4 is not so regular too.

Unlike other groups, Group 5 may be midnight travel vehicle, which has the most distinguish feature. Vehicles start their travel at 0:40 am and the activity duration is around 17.14 hours. Besides, the number of travel times per day is 2.99, which is also higher than others and the travel distance varies between 4.8 km and 32.5 km. The origins and destinations also have a certain degree of randomness.

Group 6 is defined as peak-hour travel with short activity duration vehicle. It starts the first travel at 8:55 am and finishes at 6:11 pm. They travel 1.82 days in a week and 2.71 trips in a

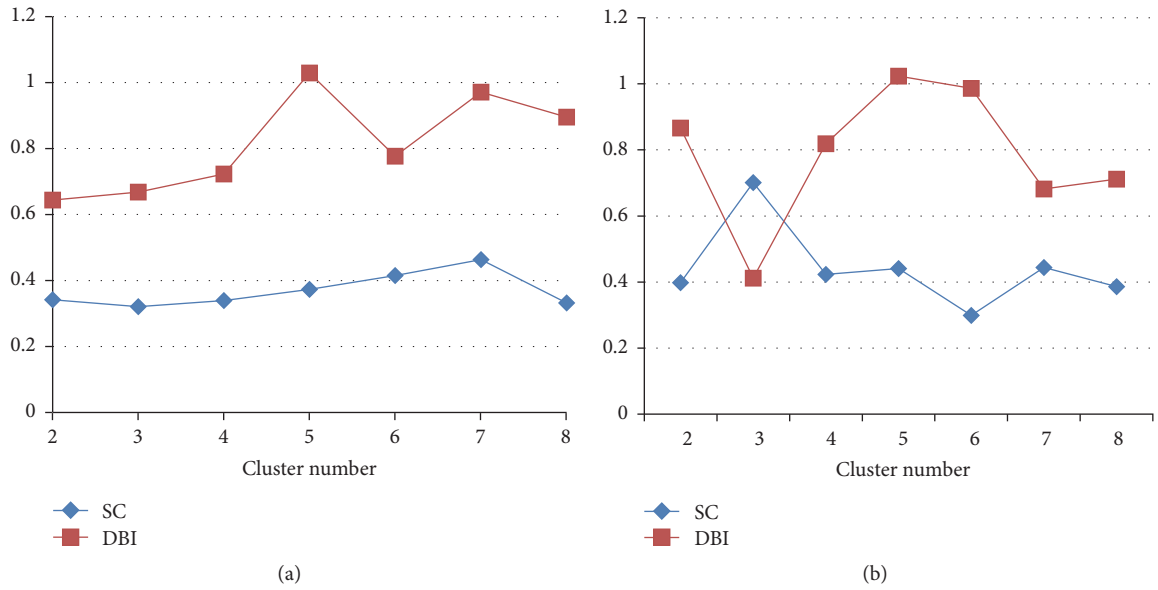

Figure 5: DBI and SC of weekdays and weekends. (a) Weekdays. (b) Weekends.

day and they have short activity duration. The travel origins and destinations are not regular and they travel for 28.1 km in average.

In general, the start time of the first trip and the end time of the last trip for Group 2 are similar to those of Group 6, both in the peak hour. Even so, the days of Group 6 traveling in a week are less and its travel distance is much longer. Comparing the characteristics of Group 2 with Group 6, we can conjecture that Group 2 is commuting vehicles traveling twice everyday and Group 6 may be vehicles commuting only in part of the days in a week and consistent with activities for leisure, recreational, or sporadic work in the rest days. Moreover, Groups 2, 3, and 4 are the main composition of traffic flow, taking up 79.1% of the whole vehicle population. Group 4 is off-peak hour travel vehicle and there is no clear travel purpose that could be inferred using only these travel behavior characteristics. These clusters could be composed of leisure travelers, visitors, or sporadic vehicles. They may be vehicles coming out to pick up child or shopping nearby. Group 5 has distinguishing features from others; they travel only in the midnight; it is similar to taxi or online hailing vehicles (i.e., Uber); the travel time,and travel purposes are random and not sure.

For Vehicles in Weekends, Three Groups Are Clustered. The characteristics of each group are shown in Table 6.

Group 1 is deduced as off-peak hour travel, where the starting time of the first travel is 10:11 am and the trip ending time is 19:30 pm. They travel 1.87 days in a week and 2.02 trips in a day. In addition, the average travel distance is about 30.8 km and the similarity of the travel origins and destinations is high. Combining with the travel frequency, travel time, and travel distance of these vehicles, they may live in the city center for work in the weekdays and during weekends they may visit their parents or relatives in the suburbs or have picnics to relax.

Group 2 is defined as afternoon travel with short activity duration vehicle, which travels 1.27 trips per day and 1.66 days in a week. It travels in off-peak hour, which is 12:30 am and 16:18 pm, the travel distance is not long and the activity duration is about 3 hours. Additionally, the origins and destinations are relatively stable. Combined with all of these features, group 2 tends to be vehicles going shopping or leisure on weekends.

Group 3 may be peak-hour travel vehicle, the average start time of the first trip is 7:42 am and the average finish time of last trip is 18:50 pm and it only travels 2.11 days in a week. The travel distance is as short as Groups 3 and 4 in weekdays. Vehicles in this group resemble commuting vehicles in weekdays. This kind of vehicles may work only in weekends, for example, people working for cram schools and the like.

6. Conclusions

This paper shows that it is possible to analyze the travel characteristics of vehicles and identify vehicle groups with similar travel behavior using LPR data. The main contribution of this paper is summarized as follows:

(i) Six vehicle groups with similar travel characteristics in weekdays and three groups in weekends are identified and the detailed behavior of each cluster is presented.

(ii) Travel characteristics are studied by analyzing the distribution of these variables and the values of each

Table 5: Average values of variables for each category in weekdays.

	Type	Days of travel per week (day)	Number of trips per day	Average start time of first trip	Average finish time of last trip	Total activity duration (h)	Maximum/minimum/average travel distance (km)	The proportion of different start/end points	Percentage (%)
(1)	Long travel distance	1.82	2.13	10:14	19:02	7.41	78.1/24.5/30.3	0.92/0.91	9.5
(2)	Commuting	**5.94**	**2.18**	**8:42**	**18:18**	**8.67**	**59.2/6.9/17.5**	**0.12/0.09**	**33.7**
(3)	Noon travel	1.08	1.82	10:07	**15:14**	4.02	3.5/1.9/2.2	0.67/0.82	29.6
(4)	Off-peak hour travel	1.82	1.63	10:20	19:48	9.09	2.9/1.2/2.4	0.77/0.73	15.8
(5)	Midnight travel	3.13	2.99	**0:40**	21:30	**17.14**	32.5/4.8/14.3	0.66/0.57	7.3
(6)	Peak-hour travel	1.82	2.71	**8:55**	**18:11**	9.07	68.0/12.2/28.1	0.97/0.98	4.1

Table 6: Average values of variables for each category in weekends.

	Type	Days of travel per week (day)	Number of trips per day	Average start time of first trip	Average finish time of last trip	Total activity duration (h)	Maximum/minimum/average travel distance (km)	The proportion of different start/end points	Percentage (%)
(1)	Off-peak hour travel	1.87	2.02	**10:11**	**19:30**	12.48	46.8/21.6/30.8	0.19/0.21	28.9
(2)	Afternoon travel with short activity duration	1.66	1.27	12:30	16:18	2.98	10.2/4.8/7.9	0.34/0.27	39.5
(3)	Peak-hour travel	2.11	1.85	**7:42**	**18:50**	6.81	4.6/1.9/3.7	0.24/0.31	31.6

variable for each category. In addition, we defined vehicle type for each group of vehicle, to identify the commuting vehicle and other ordinary leisure travel vehicles, and the clustering result can be used in several aspects, such as

(1) policy making in vehicle classification management;
(2) transport planning and vehicle travel forecasting;
(3) urban traffic simulation and monitoring.

For example, with the clustering, we can effectively extract the commuting travel vehicles which provide better decision information for developing urban traffic demand and managing policy by analyzing the spatial and temporal distribution of its travel behavior. In addition, summarizing the clustering result, there are almost 46% (type 3 and type 4) off-peak hour travel vehicles traveling in short distance (less than 3.5 km) in weekdays. Considering that the detectors are mainly installed on expressways, we can guide these vehicles to take arterial roads instead of expressways by implementing some traffic management schemes during off-peak hour to improve the level of services of arterial roads and finally release the traffic pressure of off-peak hours on expressways.

In general, firstly, this study has shown that it is possible to analyze the travel characteristic of vehicles and identify vehicle groups with similar travel behavior using LPR data. Besides, a study of the vehicles' travel pattern can be performed based on this study results and this information can be used to preferably understand how the behavior of the different groups affects the road system, the travel patterns, and travel modes.

Secondly, from the standpoint of transportation planning, clustering vehicle travel patterns allow the analysis of possible differences in level of service experienced by different vehicle segments and the identification of potential biases. It can also provide better understanding of how changes in level of service affect different vehicles and how they respond to those changes. Knowing the main differences between groups can contribute to a better understanding of the effect of disruptions on travel behavior.

Finally, the method displayed in this study is innovative and practical which can be applied in several similar problems and researches. It highlights the potential of using LPR data to mine underlying information of vehicles and the study also reveals the importance of clustering vehicles based on their characteristics.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This research was sponsored by National Natural Science Foundation of China (71171147) and Fundamental Research Funds for the Central Universities.

References

[1] W. Wenjing and G. Hongcheng, "Car and rail travel mode choice behavior analysis," *Urban Transportation of China*, vol. 3, pp. 11–14, 2010.

[2] A. Chakirov and A. Erath, "Use of public transport smart card fare payment data for travel behaviour analysis in Singapore," in *Proceedings of the 16th International Conference of Hong Kong Society for Transportation Studies*, Hong Kong, 2011.

[3] L. Liu, A. Hou, A. Biderman, C. Ratti, and J. Chen, "Understanding individual and collective mobility patterns from smart card records: a case study in Shenzhen," in *Proceedings of the 12th International IEEE Conference on Intelligent Transportation Systems (ITSC '09)*, pp. 842–847, St. Louis, Mo, USA, October 2009.

[4] T. H. Rashidi, A. Abbasi, M. Maghrebi, S. Hasan, and T. S. Waller, "Exploring the capacity of social media data for modelling travel behaviour: opportunities and challenges," *Transportation Research Part C: Emerging Technologies*, vol. 75, pp. 197–211, 2017.

[5] E. Chung, "Classification of traffic pattern," in *Proceedings of the 10th World Congress on Intelligent Transport Systems*, vol. 11, pp. 16–20, Madrid, Spain, 2003.

[6] B. S. Everitt, S. Landau, and M. Leese, *Cluster Analysis*, Edward Arnold, London, UK, 2001.

[7] S. Hanson and J. Huff, "Classification issues in the analysis of complex travel behavior," *Transportation*, vol. 13, no. 3, pp. 271–293, 1986.

[8] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[9] H. Ling and W. Lingda, "Summary of clustering algorithms in data mining," *Application Research of Computers*, vol. 1, pp. 10–13, 2007.

[10] W. Weijermars and E. Van Berkum, "Analyzing highway flow patterns using cluster analysis," in *Intelligent Transportation Systems*, pp. 308–313, 2005.

[11] T. Li, T. Pei, Y. C. Yuan et al., "A summary for the classification of pattern and application of human trajectory," *Progress in Geography*, vol. 33, no. 7, pp. 938–948, 2014.

[12] B. Zhang, *Research on Taxi Trajectory Data Mining Based on Cloud Computing*, Xidian University, Xi'an, China, 2014.

[13] L. Gong, H. Sato, T. Yamamoto, T. Miwa, and T. Morikawa, "Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines," *Journal of Modern Transportation*, vol. 23, no. 3, pp. 202–213, 2015.

[14] L. Gong, H. Sato, T. Morikawa et al., "Activity stop and non-activity stop identification in GPS trajectories utilizing density-based clustering method and support vector machines," in *Proceedings of the Transportation Research Board Annual Meeting*, 2015.

[15] X. Zhang and W. Guang, "Study on urban traffic road segmentation based on cluster analysis," *Intelligent Transportation Systems and Information Technology*, vol. 9, no. 3, pp. 36–41, 2009.

[16] H. R. Arkian, R. E. Atani, and S. Kamali, "Cluster-based traffic information generalization in vehicular ad-hoc networks," in *Proceedings of the IEEE International Symposium on Telecommunications*, pp. 197–207, 2014.

[17] E. Jenelius and H. N. Koutsopoulos, "Travel time estimation for urban road networks using low frequency probe vehicle data,"

Transportation Research Part B: Methodological, vol. 53, no. 4, pp. 64–81, 2013.

[18] N. H. M. Wilson, J. Zhao, and A. Rahbee, "The potential impact of automated data collection systems on urban public transport planning," in *Schedule-Based Modeling of Transportation Networks*, vol. 46, pp. 1–5, 2009.

[19] G. Leduc, "Road traffic data: collection methods and applications," JRC Technical Notes, Working Papers on Energy, Transport and, Climate Change 1, 2008.

[20] F. Öztürk and F. Özen, "A new license plate recognition system based on probabilistic neural networks," *Procedia Technology*, vol. 1, pp. 124–128, 2012.

[21] M. A. Massoud, M. Sabee, M. Gergais, and R. Bakhit, "Automated new license plate recognition in Egypt," *Alexandria Engineering Journal*, vol. 52, no. 3, pp. 319–326, 2013.

[22] R. Camus, G. Longo, and C. Macorini, "Estimation of transit reliability level-of-service based on automatic vehicle location data," *Transportation Research Record*, vol. 1927, pp. 277–286, 2005.

[23] S. Lee and M. Hickman, "Trip purpose inference using automated fare collection data," in *Proceedings of the 4th Transportation Research Board Conference on Innovations in Travel Modeling*, Tampa, Fla, USA, 2012.

[24] X. Zhan, R. Li, and S. V. Ukkusuri, "Lane-based real-time queue length estimation using license plate recognition data," *Transportation Research Part C: Emerging Technologies*, vol. 57, pp. 85–102, 2015.

[25] C. Hong, X. Xiangchun, Y. Feng, and J. Zhou, "A novel method of trip route estimation based on vehicle license plate recognition system," in *Proceedings of the 13th COTA International Conference of Transportation Professionals (CICTP '13)*, November 2013.

[26] M. P. Dixon and L. R. Rilett, "Population origin-destination estimation using automatic vehicle identification and volume data," *Journal of Transportation Engineering*, vol. 131, no. 2, pp. 75–82, 2005.

[27] M. A. Munizaga and C. Palma, "Estimation of a disaggregate multimodal public transport Origin-Destination matrix from passive smartcard data from Santiago, Chile," *Transportation Research Part C: Emerging Technologies*, vol. 24, pp. 9–18, 2012.

[28] H. Zhang, X. Liu, X. Duan, D. Miao, and H. Ma, "A comparative study of clustering algorithms in data mining," *Computer Applications and Software*, vol. 20, no. 2, pp. 5–6, 2003.

[29] A. Al-Wakeel and J. Wu, "K-means based cluster analysis of residential smart meter measurements," *Energy Procedia*, vol. 88, pp. 754–760, 2016.

[30] A. E. Raftery and N. Dean, "Variable selection for model-based clustering," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 168–178, 2006.

[31] J. Yuqing and G. Dunwei, "Fast genetic clustering algorithm," in *Proceedings of the China Intelligent Automation Conference*, vol. 38, pp. 186–190, 2007.

[32] L. Zhu, B. Ma, and X. Zhao, "Effectiveness analysis based on clustering coefficient contour," *Journal of Computer Applications*, vol. 30, pp. 139–141, 2010.

[33] H. Cai, X. Zhan, J. Zhu, X. Jia, A. S. F. Chiu, and M. Xu, "Understanding taxi travel patterns," *Physica A: Statistical Mechanics and Its Applications*, vol. 457, pp. 590–597, 2016.

[34] C. Fraley and A. E. Raftery, "How many clusters? Which clustering method? Answers via model-based cluster analysis," *The Computer Journal*, vol. 41, no. 8, pp. 578–588, 1998.

[35] S. Jigui and J. Jie, "Research on clustering algorithm," *Journal of Software*, vol. 19, pp. 48–61, 2009.

[36] H. Nishiuchi, J. King, and T. Todoroki, "Spatial-temporal daily frequent trip pattern of public transport passengers using smart card data," *International Journal of Intelligent Transportation Systems Research*, vol. 11, no. 1, pp. 1–10, 2013.

[37] M. A. Ortega-Tong, *Classification of London's public transport users US smart card data [Bachelor of Science in Civil Engineering]*, University of Chile, 2007.

[38] P. Jones and M. Clarke, "The significance and measurement of variability in travel behaviour," *Transportation*, vol. 15, no. 1-2, pp. 65–87, 1988.